(3)

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 March 2001 (08.03.2001)

PCT

(10) International Publication Number
WO 01/16810 A2

(51) International Patent Classification[7]: G06F 17/50

(21) International Application Number: PCT/EP00/08504

(22) International Filing Date: 31 August 2000 (31.08.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/387,741 31 August 1999 (31.08.1999) US

(71) Applicant *(for all designated States except US)*: **THE EUROPEAN MOLECULAR BIOLOGY LABORATORY** [DE/DE]; Meyerhofstrasse 1, 69117 Heidelberg (DE).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **LACROIX, Emmanuel** [BE/DE]; Im Schilling #24, 69181 Leimen (DE). **SERRANO, Luis** [ES/DE]; Zaerringerstrasse 19, 69117 Heidelberg (DE).

(74) Agents: **LEGG, Cyrus, James, Grahame** et al.; Abel & Imray, 20 Red Lion Street, London WC1R 4PQ (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/16810 A2

(54) Title: A COMPUTER-BASED METHOD FOR MACROMOLECULAR ENGINEERING AND DESIGN

(57) Abstract: The present invention relates to a system and method for engineering and designing a macromolecule. An experimentally determined or *de novo* atomic structure that corresponds to the macromolecule is identified. The atomic structure is composed of building blocks. When the macromolecule is a peptide or a protein, the building blocks are amino acid residues. A target subset of the building blocks in the atomic structure to be optimized is identified. The coordinates of those building blocks that are not in the target subset are fixed. For each building block in the target subset, a large number of potential conformers is sampled. Each conformer to be sampled is substituted into the atomic structure and tested against an energy function that includes the equivalent energy of the conformer in a reference state and comprises at least one entropic term. Combinations of conformers that best satisfy an interaction energy function are identified.

A COMPUTER-BASED METHOD FOR MACROMOLECULAR ENGINEERING AND DESIGN

1. FIELD OF THE INVENTION

The present invention relates to methods for engineering and designing molecules which comprise building blocks that are individually amenable to systematic variation. Particular areas of application include the design and development of macromolecules, for example, proteins, peptides, nucleic acids and polymers with desired properties such as stability and specificity of interaction with counterpart molecules. Specifically, the present invention relates to computer-based methods that employ search methods in the space of available molecules or fragments thereof which could form building blocks of a molecular structure and which use a three dimensional description of the structure with atomic scale resolution. An aim of the invention is to provide guidance to the experimental scientist who is not able to systematically consider all of the possible combinations of building blocks which might need to be permuted.

2. BACKGROUND OF THE INVENTION

Biochemistry and synthetic chemistry are replete with molecules whose structures consist of hundreds or thousands of atoms but whose constitution can also be thought of as that of a sequence of identifiable units, each of which comprises only a small number of atoms. Molecules of this sort will herein be referred to as macromolecules, a term which can be taken to include proteins, peptides, cyclic peptides, nucleic acids, lipids, carbohydrates and synthetic polymers, etc. The individual units which go to make them up will be called building blocks, though other terms, both generic and specific, may be found in the art. For example, the building blocks of proteins and peptides are amino acid residues, the building blocks of nucleic acids are nucleotides, and synthetic polymers are built from monomers. The term monomer can also be used in a more general sense. A building block, on its own, will usually differ slightly from its bound form in the macromolecule: the reaction with the building blocks which become its neighbors in the macromolecule structure may truncate its own structure. In this way a different term is often used for the free building block from its bound form. For example, amino acids are the building blocks of peptides and proteins whereas in their bound form they are called residues. The term building block, as used herein, will be understood to mean both the free molecule and its bound form, unless otherwise evident from the context in which it is used. The chemistry and other properties of macromolecules are often understood in terms of the

types of building blocks employed and the order in which they are found, i.e., their sequence.

Protein or peptide engineering is a process using recombinant DNA technology or chemical methods to modify the amino acid sequences of natural proteins or peptides to improve or alter their function. By changing, i.e., mutating, the natural amino acid sequence of a protein or peptide, it is possible to alter, *inter alia*, its stability, substrate specificity, activity, and inter- and intra-molecular interactions. Changes to an amino acid sequence can be made on a purely random basis, or can be derived from educated guesses based on the atomic-resolution detail of the protein or peptide three dimensional structure provided by techniques such as X-ray crystallography, nuclear magnetic resonance, electron microscopy, and electron crystallography.

The mutagenesis of proteins and peptides, even when carried out non-randomly using structural information, can have unexpected or undesired results. First, a mutant protein or peptide may not have any altered characteristics from its native counterpart. Second, a mutant protein or peptide may acquire completely different characteristics from those desired. Third, a mutant protein or peptide may not be properly folded, rendering it unstable, insoluble, lethal, or completely non-functional. Protein engineering can thus be a trial-and-error process for generating a properly folded mutant protein or peptide with a desired function. Furthermore, mutagenesis to probe the function of proteins, so-called "site-directed mutagenesis", is a time-consuming process, involving introduction of the mutation into the DNA coding region, transformation of the mutated sequence into the appropriate cells, expression of the protein, purification, and functional assays.

Often, desired changes in protein or peptide function, *e.g.*, altered binding specificity or avidity, require the simultaneous mutagenesis of several amino acids. With twenty possible naturally occurring amino acids at each position, the number of variants that need to be screened is enormous. For changes at three amino acids, there are 8,000 possible combinations; for changes at 10 amino acids, 10^{13} different amino acid sequences are possible. Even though the number of variants may be narrowed by making educated guesses based on knowledge of the protein or peptide structure, a large number of mutants may still have to be made in order to engineer a properly folded protein or peptide with the desired characteristics.

Ab initio peptide and protein design presents more difficulties than the engineering of mutant proteins and peptides. In *ab initio* design, nearly *all* of the amino acids must be chosen to create a properly folded peptide or protein with a desired function, making the number of possible variants even greater than for conventional mutagenesis. Furthermore, if the fold of the functional protein or peptide is not well-characterized, or if the structure

cannot be designed based on the known structure of an homologous protein (homology modeling), then structural information will not be available to help narrow down those combinations of amino acids that are most likely to adopt the proper protein fold. Therefore, *ab initio* design of proteins and peptides by *in vitro* production and testing of all amino acid sequence variants is impractical, if not impossible.

Computer-based methods for designing and engineering proteins and peptides should allow for the identification of amino acid sequence variants that can be accommodated by the three dimensional structure of the protein being mutated, thus decreasing the number of *in vitro* engineering experiments that need to be performed. The central principle is that it is far easier to consider a large number of sequence variations and choose the best candidates through computer simulation than it is through direct experimentation and synthesis in the laboratory.

Nevertheless, computational methods are non-trivial because of the complexity of the problem and the quality of the primary data that is accessible for immediate use. For example, the high-resolution three dimensional structures of most small organic molecules are available, but those of proteins are typically at lower resolution. Further, the methods of modeling the weak, non-covalent forces, *e.g.*, hydrogen bonds, van der Waals interactions, and hydrophobic interactions, that maintain the three-dimensional structures of macromolecules are at present very crude. And, in general, the number of degrees of conformational freedom that are required to accurately describe the structure of a protein is too large to enable practical exploration of its potential energy surface. Our ability to reliably model small changes in a protein structure is therefore limited by several factors: the accuracy to which the whole structure is known; the impracticality of applying usual optimization methods to systems as large and complicated as proteins; and the inaccuracy of the intermolecular potential functions which are needed to model the ways in which residue side chains determine the three dimensional structure of the protein by aligning with one another.

For these reasons, current computer-based methods for designing and engineering macromolecules cannot efficiently and reliably predict the accommodation of variant structures by an identified protein fold, and thus have limited utility in assessing which sequence variants are likely to have a desired structure and function.

Previous computational approaches to protein engineering have been limited to predictions of tertiary structure from sequence, geometric rather than energetic positioning of side chain atoms, and prediction of favorable sites of cross-linking.

In an analogous way, the exploration of nucleic acid structures is subject to the same complexities of mathematical modeling, as well as the combinatorial problem arising from

the fact that at least 4 different nucleotides can be considered for each position on the sequence.

Clearly, there is a need for computer-based methods for designing and engineering macromolecules which utilize a more complete and accurate mathematical description of macromolecular structure than has hitherto been attempted but which employ practical and reasonable approximations to enable efficient execution. In this way it will be possible to predict the structures of macromolecular variants and enable the selection of variants having a desired structure and function.

Citation of a reference herein shall not be construed as indicating that such reference is prior art to the present invention.

3. SUMMARY OF THE INVENTION

The present invention relates to an improved computer-based method for optimizing specific building blocks in the sequence set of building blocks which make up a target macromolecule, for example the amino acid residues of a peptide or protein. In essence, the central features of the invention are, given a set of substitute building blocks and a set of positions in the sequence of the target macromolecule, use of a plurality of conformations or conformers of each building block; use of a scoring function to quantify and rank the possible structures; use of a reference structure; and use of filtering techniques for simplifying the analysis.

In detail, the method comprises the steps of: (a) specifying at least one substitute building block for each position in a set of positions; (b) determining, for each substitute building block, one or more candidate conformers and: (i) substituting the coordinates of each candidate conformer or portion thereof for the corresponding building block or portion thereof in a structure of atomic resolution of said target macromolecule; and (ii) calculating an intrinsic energy of each candidate conformer; (c) rejecting candidate conformers having an intrinsic energy above a threshold value and whose weights, computed by a partition function are below a threshold value; (d) calculating a pairwise interaction energy for all possible pairs of candidate conformers that have not been rejected in step (c) and combining the sum of the pairwise interaction energies for all pairs with the sum of the intrinsic energies for all candidate conformers to give a solution score; (e) determining solution structures, from a plurality of solution structures, which have, respectively, solution scores that comprise an entropic term and that are lower than a predetermined threshold solution score; wherein: (i) each building block in said building block set is represented in each solution in said plurality of solutions by one or more candidate conformers that each correspond to a substitute building block that was independently specified in accordance

with step (a) and was not rejected in step (c); and (ii) each solution score representing a difference in the summed potential energy of each candidate conformer in the solution structure when a candidate conformer is substituted in an atomic-scale resolution structure of the target macromolecule, and when said candidate conformer is substituted into an atomic resolution macromolecular structure corresponding to a reference structure. The application of the method may be carried out more than once, sequentially, to obtain better and better solutions. The solutions may then be used as suggestions for synthetic candidates and those molecules which are made may then be assayed against a target of interest.

The present invention further relates to a computer program product for use in conjunction with a computer, the computer program mechanism comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising an application program configured to choose a set of substitute building blocks in a target macromolecule, the application program, (a) upon receiving a request to choose a set of building blocks for a set of positions in the sequence of the target macromolecule and, being input at least one substitute building block for each position in said set of positions; (b) determining, for each substitute building block, one or more candidate conformers and: (i) substituting the coordinates of each candidate conformer or portion thereof for the coordinates of the corresponding building block or portion thereof in a structure of atomic resolution of said target macromolecule; and (ii) calculating an intrinsic energy, of the candidate conformer; (c) rejecting candidate conformers having an intrinsic energy above a threshold value and whose weights, computed by a partition function are below a threshold value; (d) calculating a pairwise interaction energy, for all possible candidate conformers that have not been rejected in step (c) and combining the sum of the pairwise interaction energies for all pairs with the sum of the intrinsic energies for all candidate conformers to give a solution score; (e) determining solution structures, from a plurality of solution structures, which have, respectively, solution scores that comprise an entropic term and that are lower than a predetermined threshold solution score; wherein: (i) each building block in said building block set is represented in each solution in the plurality of solutions by one or more candidate conformers that each correspond to a substitute building block that was independently specified in accordance with step (a) and was not rejected in step (c); and (ii) each solution score representing a difference in the summed potential energy of each candidate conformer in said solution structure when a candidate conformer is substituted in an atomic-scale resolution structure of the target macromolecule, and when the candidate conformer is substituted into an atomic resolution macromolecular structure corresponding to a reference structure.

The present invention further comprises a system for choosing a set of building blocks in a target macromolecule comprising: a central processing unit; an input device for inputting requests; an output device; a memory; at least one bus connecting the central processing unit, and the input device; the memory storing an application program configured to choose a set of substitute building blocks in the target macromolecule, the application program, (a) upon receiving a request to choose a set of building blocks for a set of positions in the sequence of the target macromolecule and being input at least one substitute building block for each position in the set of positions; (b) determining, for each substitute building block, one or more candidate conformers and: (i) substituting the coordinates of the candidate conformer or portion thereof for the coordinates of the corresponding building block or portion thereof in a structure of atomic resolution of said target macromolecule; and (ii) calculating an intrinsic energy, of the candidate conformer; (c) rejecting candidate conformers having an intrinsic energy above a threshold value and whose weights, computed by a partition function are below a threshold value; (d) calculating a pairwise interaction energy, over all possible candidate conformers that have not been rejected in step (c) and combining the sum of the pairwise interaction energies for all pairs with the sum of the intrinsic energies for all candidate conformers to give a solution score; (e) determining solution structures, from a plurality of solution structures, which have, respectively, solution scores that comprise an entropic term and that are lower than a predetermined threshold solution score; wherein: (i) each building block in said building block set is represented in each solution in the plurality of solutions by one or more candidate conformers that each correspond to a substitute building block that was independently specified in accordance with step (a) and was not rejected in step (c); and (ii) each solution score representing a difference in the summed potential energy of each candidate conformer in said solution structure when a candidate conformer is substituted in an atomic-scale resolution structure of the target macromolecule, and when the candidate conformer is substituted into an atomic resolution macromolecular structure corresponding to a reference structure.

4. BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1: Block diagram of a computer system in accordance with the present invention.

Figure 2: Flow chart describing the steps of data input and structure validation.

Figure 3: Flow chart describing steps of intrinsic energy computation and rejection of rotamers.

Figure 4: Flow chart describing steps of pairwise energy computation and selection of solution structures in order to compute solution score.

Figure 5: Flow chart describing steps of pairwise energy computation and selection of solution structures by use of a genetic algorithm in order to compute solution score.

Figure 6: Van der Waals interaction energy for two methyl ($-CH_3$) groups as a function of the distance that separates them.

Figure 7: Electrostatic interaction between two unit charges of equal sign, either according to Equations 26 and 27 and using a relative dielectric constant E_r of 4.0 r_{ij} (solid line), or according to Equation 28 using the same dielectric constant and a screening distance r_s of 2.0 (dashed line).

Figure 8: Geometrical conditions to be satisfied in hydrogen bonding.

Figure 9: Schematic representation of the accessible surface area (ASA) of a protein. This surface is defined as the surface marked by the center of a water molecule (a probe P with radius 1.4 Å) rolling around the protein while maintaining permanent contact with the van der Waals surface of the protein atoms. The arrows indicate that the solvation potential of carbon and oxygen atoms correspond to positive and negative energies, respectively, so that carbon atoms tend to cluster inside the protein while oxygen atoms prefer to protrude outside. The bold surface close to the C and O atoms are their atomic accessible surface areas.

Figure 10: Conformers generally observed around a rotatable single bond (from left to right: gauche -, trans and gauche +). X and Y represent two heavy atoms, *e.g.*, the alpha and delta carbons of a leucine side chain.

Figure 11: Distribution of χ_1 (Val) and χ_1, χ_2 (Leu) dihedral angle values using Gaussian equations (black). Dark grey curves represent the distribution of the trans, gauche - and gauche + conformers. In light grey are the distributions corresponding to non-rotameric configurations.

Figure 12: Illustration of the rotamer library concept, showing the side chain conformers of valine and leucine. Numbers on top of each structure are the χ_1 (Val) and χ_1, χ_2 (Leu)

dihedral angle values; those labelled with an asterisk were obtained during the evaluation of *Perla*.

Figure 13: Accompanying example 1, distribution of solution scores for 1600 output sequences from *Perla*, applied to the SH3 domain of alpha-spectrin. Wild Type sequence is indicated along with score of best solution and worst solutions found.

Figure 14. Far-UV CD spectra of the SH3 protein domains, at pH 3.5: (A) wild-type protein, (B) first core mutant and (C) second core mutant. (Empty circles) 278K. (Filled circles) 298K. (Filled squares) 363K.

Figure 15. Far-UV CD spectra of the SH3 protein domains, at pH 3.5: (A to C) RT-loop, diverging turn and distal loop protein mutants, respectively. (Empty circles) 278K. (Filled circles) 298K.

Figure 16. Urea-induced denaturation of the SH3 domain proteins, monitored through emission fluorescence (total fluorescence above 340nm, excitation wavelength was 280nm). Experimental conditions were pH3.5 and 298K. (Circles) Wild-type protein, (squares) first core mutant and (triangles) second core mutant. Filled and empty symbols represent the data obtained from two separate experiments. The lines are the results of a least-square regression fitting of the data to Equation 54.

Figure 17. Urea-induced denaturation of the SH3 domain proteins, monitored through emission fluorescence (total fluorescence above 340nm, excitation wavelength was 280nm). Experimental conditions were pH3.5 and 298K. (Circles) Wild-type protein, (squares) RT-loop mutant and (triangles) distal loop mutant. Filled and empty symbols represent the data obtained from two separate experiments. The lines are the results of a least-square regression fitting of the data to Equation 54.

5. DETAILED DESCRIPTION OF THE INVENTION

Section 5.1 gives an overview of the invention including a computer system for optimizing a set of building blocks in a macromolecule. In Section 5.2, the method of the present invention as implemented in *Perla*, the preferred embodiment of the invention, is described in brief. Subsequent sections describe in more detail each step of the method of the present invention, with emphasis on these steps as implemented in *Perla*. In Section

5.3, a detailed mathematical description is given of the empirical scoring function used to calculate the energy difference between an optimized conformer of a mutated target protein and some reference state. Section 5.4 describes a modified form of the scoring function that is split into template, intrinsic and pairwise terms. Section 5.5 provides a detailed theoretical description of the molecular mechanics potential, and of van der Waals, electrostatic, and hydrogen bonding energies, which contribute to it. Section 5.6 provides a detailed mathematical description of the empirical potential, calculated from changes in solvation and entropy of the protein chain, and which introduces an approximate description of the interaction of solvent with the side chain conformers. Section 5.7 describes how the denatured state of proteins are considered in *Perla*. Section 5.8 describes the generation of the rotamer library used by *Perla*. Section 5.9 describes energy optimization and elimination of incompatible amino acid conformers. Optimization routines, *e.g.*, dead-end elimination and mean field theory, are detailed in Section 5.10. Section 5.11 describes re-evaluation of solvation energies, and consequently, of the scoring function, for sequences that remain after elimination, and Section 5.12 details output from the preferred embodiment of the present invention. Section 5.13 details a generalization of the method to macromolecules other than proteins and peptides.

5.1. OVERVIEW

The present invention relates to a novel method for designing and engineering macromolecules that utilizes an accurate and complete mathematical representation of macromolecular structure, in order to reliably predict how precise variants of its sequence can be accommodated into a desired three-dimensional (3D) structure. Herein, the 3D structure in question may be a specific conformation of the macromolecule itself or may be a complex in which the macromolecule interacts with a ligand or another macromolecule. A preferred embodiment of the present invention is known as *Perla*. *Perla* is a computer-based method for protein engineering that manipulates peptides and proteins in order to identify and sort amino acid sequences capable of folding into a desired 3D structure.

In order to model the effect of a small change on a protein structure, it is not always necessary to reanalyze its entire structure. Consequently, well chosen detailed structural information about the site of mutation may be employed to focus attention on the area of interest. Structural flexibility of a protein may be thought of as a large-scale consequence of the conformational flexibility of the building blocks of which it is composed. Here we may exploit the fact that residue mutation in a protein is effectively a side chain substitution which leaves the backbone unperturbed. That is, the part of the structure of the amino acid building block which changes from one to another is the side chain alone. Correspondingly,

conformational analysis may be simplified by using sets of known favorable side chain conformations instead of carrying out an unconstrained energy minimization.

To identify sequences that are most able to fold into the target protein 3D structure, an energy or scoring function is established, referring to a conformational state representing either the denatured or unfolded protein. A random coil or extended structure can also be used as reference. The scoring function, as a difference between an energy of the protein folded structure, and an energy of the protein denatured or unfolded state is correlated to the stability of the folded protein structure.

Although the invention is described herein below mainly with application to proteins and peptides, as will be evident to a skilled artisan, the teaching herein can be readily adapted for use with other macromolecules such as nucleic acids, carbohydrates and other polymers.

5.1.1. A SYSTEM FOR OPTIMIZING STRUCTURAL UNITS OF A MACROMOLECULE

The present invention addresses the ability to engineer derivatives of large molecules through systematic variations of their structural components by presenting solution scores of preferred solution structures after rigorous analysis of a user-defined search space. The invention, as shown in figure 1, comprises a system 100 for optimizing a set of structural units in a target macromolecule, comprising at least one central processing unit 102, a user interface 104 for inputting requests, an output device 106, a section of main memory 110, and a bus 108 connecting the central processing unit, the memory, the input device, and the output device. Optionally, the system is connected to a network interface 114, via which access to a Protein Database 116, such as the PDB, can be achieved. The memory stores an operating system, 120, a file system 122, at least one set of molecular mechanics parameters 124, a cache 126 and an Application program 130, also known as *Perla*. In a preferred embodiment, system 100 comprises an multi-processor computer, capable of carrying out calculations in parallel.

Application program 130 is configured to optimize a set of structural units in the target macromolecule and comprises a number of modules, including but not limited to: a solution score module 132, to compute a solution score of a mutated or target structure; a sub-rotamer module 134 to calculate distributions of sub-rotamers for rotamer side chains; an input reader 136; optionally, a library of pre-computed rotamer conformations 138; a solvation energy module 140 including functions that assign residues to be "core" and "non-core"; a surface area module 142 for computing the solvent accessible area or other surface of a target structure or mutated form thereof; a dihedral angle module 144 for computing

backbone dihedral angles on the target structure or mutated form thereof; a penalty function module 146 for computing a statistically derived penalty function; a mean field theory module 148 for optimizing side chain distributions and for computing side chain contributions to the entropy; a genetic algorithm module 150 for selecting an optimum family of mutated sequences and/or conformers; and a dead end elimination module 152 for eliminating unfavorable sequences and/or conformers; and a collection of protein fragments, 154. Additional modules, not shown in Figure 1 may comprise a module to assign atom types to atoms in an input structure, in order to assign molecular mechanics force field parameters and modules containing geometric manipulation utilities, such as may be required to translate rotamer coordinates from a library to the template.

The macromolecule, which may be a peptide, protein, strand of DNA or RNA or a carbohydrate, or any organic molecule which consists of identifiable distinct structural units, has a 3D structure whose geometry is known to atomic resolution. The Application Program 130, upon receiving a request to choose a set of substitute building blocks for at least one set of positions, utilizes, for each substitute building block, one or more candidate conformers. For each determined candidate conformer, the application program substitutes a building block at a position in the target macromolecule with the candidate conformer and calculates an intrinsic energy term for the candidate conformers. The application program subsequently rejects candidate conformers having an intrinsic energy above a threshold value and according to whether the statistical weight of the conformer, calculated from a partition function, is below a threshold value. The application program calculates a pairwise interaction energy term for all possible conformers that have not been rejected by the threshold value criteria. The method enables determination of solution structures, that are ranked by solution score. In some embodiments, for example those that utilize dead end elimination, the solutions include a best solution corresponding to a global minimum energy conformation. Each building block in the set to be optimized is represented in each solution by one or more candidate conformers that were not rejected by the threshold criteria when substituted into the structure of the target macromolecule. Each solution score can be expressed as difference between the summed potential energy of each candidate conformer substituted into the target structure and the same conformer substituted into a reference structure. The solution score comprises molecular mechanics energy terms (van der Waals, hydrogen bonding and electrostatic) and terms corresponding to an empirical potential (entropy and solvation) along with a user-defined statistical term.

This system, when operated in a laboratory environment can provide an efficient and useful method of directing experimental efforts towards engineering sequence variations in a target macromolecule. Said system, being capable of quantifying the potency of a plurality of sequences and thereby selecting a small number which would be worthy of synthesis, can operate in tandem with experiment to optimize properties of interest of the target macromolecule.

5.1.2. THE PROTEIN STRUCTURE

The computer-based method of the present invention uses an "inverse folding" approach, *i.e.*, a protein backbone is chosen *a priori* as the native state to be designed and is kept fixed throughout the calculation. Fixed protein backbone atoms are the central carbon atom, Cα, and the amide group, C(=O)NH, of each residue. There is no restriction that the protein should consist of a single contiguous backbone; *Perla* can accept multiple backbones as input. The choice of a protein topology depends on the application of the engineered protein. Due to the absence of backbone motions during the evaluation of protein sequences, it is preferable for the main chain target conformation to be correctly constructed from the start. In a preferred embodiment, a protein with a well characterized protein fold or high resolution three dimensional structure is chosen (*e.g.*, from amongst those found in the Protein Data Bank (PDB), available from Research Collaboratory for Structural Bioinformatics (RCSB), web site address http://www.rcsb.org/pdb/). As used herein, the "resolution" of a macromolecular three-dimensional structure is the minimum separation two atoms can have and still appear to be distinct and separate. Thus, the higher the resolution, *i.e.* the smaller the separation distance at which two atoms can be distinguished, the more accurately determined is the structure. In a preferred embodiment, the protein model is solved at atom level resolution around the site of interest and the fixed backbone has been refined to eliminate steric clashes and unfavorable main chain dihedral angles. For this purpose the structure may have been obtained from X-ray crystallography or from NMR studies. In general, the parameters employed by the user of the invention may be chosen to best suit the quality of the data. In a second embodiment, *e.g.*, *de novo* protein design, the protein structure is not available or the protein fold is not well characterized, and methods for the construction of novel protein backbones are employed (*e.g.*, WHAT_IF; Vriend, 1990, *J. Mol. Graphics* 8:52-57; INSIGHT; Abagyan *et al.*, 1994, *J. Comp. Chem.* 15:488-506).

The 3D structures of other macromolecules can similarly be obtained from X-ray crystallography or may themselves be the outcome of mathematical or computational simulation.

5.1.3. THE AMINO ACID SIDE CHAINS

The computer-based method of the present invention uses a three-dimensional atomic description of the system to be engineered. The main chain atomic configuration being provided, the method is used to reconstruct amino acid side chains. The side chains of the twenty naturally occurring amino acids are bound to the backbone Cα atoms.

A custom-made library of discrete side chain conformations ("rotamers") for each amino acid, compiled using dihedral angle (χ_1, χ_2, χ_3, χ_4) data from available structures (preferably from those deposited in the PDB), is employed by the method of the present invention. The library of amino acid side chain conformations is preferably made by fitting occurrences of side chain dihedral angles for each amino acid side chain in known protein structures to Gaussian distributions. Since stereochemical rules were not used to generate the library, it contains side chain conformations that are not very abundant but are nonetheless important components of protein structure. Furthermore, because each dihedral angle is described by a Gaussian distribution, the observed range of oscillation of each angle is also incorporated into the library.

In application to polymeric structures other than proteins, it may be possible to derive conformer libraries from means other than by direct comparison with crystal structures. For example, stereochemical rules may be adequate for the hydroxyl groups of sugar molecules; computer simulation may be most appropriate for the modeling of nucleotide conformations. In some circumstances, the building blocks may have insufficient conformational flexibility to demand construction of conformer libraries. In such cases, the application of the method is a lot more straightforward than described herein, there being fewer conformers per building block.

5.1.4. SELECTING POSSIBLE SEQUENCES

The computer-based method of the present invention executes successive trials to consider the immense variety of sequences that can be generated as a result of protein mutagenesis, *i.e.*, substitution of one amino acid side chain with a different amino acid side chain at a given site in the protein.

In one embodiment, the user specifies which residues in the protein are to be altered. These may be specified by position in the sequence of the protein. To achieve this the user may employ specific knowledge about the 3D structure of the protein. For example, the user may choose residues which seem to be critical to folding or which are important in the definition of a binding site. In a preferred embodiment, *Perla* itself, through use of its own scoring function (see below) may automatically identify the building blocks which are to be varied. In either case, it is not necessary that the selected residues form a contiguous stretch of the sequence of the target macromolecule; nor is it necessary that any pair of the selected residues is adjacent in the sequence.

In one embodiment, the user may also specify a list of possible mutations i.e., residues to be considered at each residue position in the protein or a broad category of desirable mutations. In another embodiment, *Perla* may analyze the immediate

environment of the selected building blocks and choose mutations which are likely to cause the least disruption to that locale. For example, it may be appropriate to consider only "polar" amino acids at a particular position which is already occupied by a polar sidechain.

Sequence sampling as embodied in the method of the present invention consists of searching the required amino acid side chains within the rotamer library and fitting these onto the chosen backbone. Side chains of the amino acid residues that are not mutated can remain structurally fixed or be moved, as desired by the user of the method.

5.1.5. SCORING THE SOLUTION STRUCTURES

In order to evaluate the degree of fit of a combination of amino acid side chain rotamers to a protein structure, the method of the present invention utilizes a scoring function made up of a sum of terms which gives rise to a solution score. Unlike previous methods for protein modeling, the method of the present invention not only considers the global sum of these terms, but also requires that individual terms satisfy constraints found in natural proteins.

Because the nature of the application of the method is to produce a number of different structures, each of which is distinct from the target, it is not possible to meaningfully compare the energies of each. Consequently, the use of a reference structure for each separate solution structure enables the structures to be compared in terms of their inherent stabilities. In a preferred embodiment, in order to assign a solution score to a particular solution structure, the method calculates the difference in potential energy between the mutated protein structure and a reference structure whose sequence is the same.

A preferred embodiment of the present invention calculates a potential energy for the native and denatured (reference) states. For the latter, in a preferred embodiment, sample conformations are taken from structures present in the PDB; this method is described in more detail later. The energy difference between the two states serves as a score, and the higher the score, *i.e.*, the larger the energy difference between the two states, the better the degree of fit of the chosen sequence to the overall native-state protein structure.

The estimation of the native state potential energy requires that the optimal association of amino acid rotamers be found. For peptides longer than a few residues, an exhaustive sampling of every possible combination of rotamers is not practical. Choosing the most likely organization of side chains is a significant combinatorial problem and, therefore, the method of the present invention employs optimization routines. The underlying principle of available optimization methods, *e.g.*, dead-end elimination and

mean field theory, is that the energy is expressible as a scoring function comprising a term to describe the fixed template, one sum of terms intrinsic to every single amino acid of the sequence and a second sum for all pairs of residues.

In the preferred embodiment of the present invention, a user-defined set of rotatable side chains is modeled in the context of a fixed collection of atoms, which include main chain atoms and the side chain atoms of residues that are not included in the modeling set. Together, the fixed atoms are the template, the structure which is the direct environment of the side chains that are subject to modeling. The calculation of the sequence-independent, constant energy term corresponding to the template is not required for the evaluation of the optimal set of side chains, but can be determined in order to estimate the quality of the template structure itself. Both the intrinsic and pairwise energy terms are similar in nature and are established to correlate with observed structural parameters in proteins. The intrinsic energy term arises from interactions between the (fixed) template and the (rotatable) side chains, while the pairwise energy term arises from interactions of the (rotatable) side chains amongst themselves. Additionally, both the intrinsic and pairwise energy terms contain contributions which depend only on the nature of the residue. A van der Waals component associated with the packing of atoms, an electrostatic term associated with ion pairs, and a hydrogen bonding term, are contained in both the intrinsic energy term and the pairwise energy term of the scoring function. In the preferred embodiment of the present invention, not only the global sum of the scoring function that is considered, but also each individual term must satisfy determined constraints, as reflected in naturally occurring protein structures.

In other embodiments of the present invention and in application to macromolecules other than proteins, it may be preferable to use more than one reference state for the calculation of the scoring function. Alternatively, in other embodiments, the use of a reference state may be unnecessary.

In yet other embodiments, the scoring function may comprise a term quantifying the interaction between the macromolecule and some binding partner. For example, the macromolecule may be an enzyme and the partner its substrate; in another example, the macromolecule may be a peptide sequence and its partner may be another peptide sequence; in a further example, the macromolecule may be a nucleic acid and its partner may be a protein or some fragment thereof.

5.1.6 OPTIMIZING OR EVALUATING THE SOLUTIONS

The combinatorial problem of side chain building is solved in the method of the present invention by calculation of mean field energies. This novel integration of Mean

Field Theory, an iterative approach, into a protein modeling method provides a measure of the entropy of the molecule and allows for the consideration of all possible amino acid side chain conformers rather than just the global energy minimum, which is a more accurate description of macromolecular structure.

The foregoing steps are applicable to each individual substituted residue; the problem of considering many alternative substitute residues at many different sites is also combinatorial in scope. The invention addresses this aspect by the technique of "dead end elimination" in which certain candidate rotamers may be eliminated from the search space if their energy scores obey certain inequalities with respect to the scores of the other rotamers present in the same solution. Consequently the overall invention comprises two distinct methods of addressing problems of a combinatorial complexity.

5.2. THE STEPS OF SEQUENCE MODELING AND EVALUATION USING *PERLA*

As described in Figures 2, 3 and 4, *Perla*, the preferred embodiment of the present invention, first reads the user-specified input at step 228. The input comprises at least four pieces of information. The first is the protein structure 222, i.e., the atoms comprising the specified template (or "target") protein conformation and their Cartesian coordinates. These coordinates may have originated as fractional coordinates from a Protein Data Bank (PDB) file. There is no restriction that the atoms comprising the template form a connected unit, i.e., the protein may have multiple backbones, or several discrete proteins or peptides in juxtaposition may constitute the template. It is preferred to utilize other computational tools which ascertain the appropriate protonation state of the residues and ionize them as applicable, before passing the coordinates to *Perla*. A second item of user-specified input is a selection of amino acids 214 to engineer. A third item of user-specified input is a list of positions 210, or an indication that *Perla* should make some determination of its own in this regard. In one embodiment, the user can specify that the side chains of just those residues in the set are subject to optimization. In another embodiment, *Perla* can automatically determine which residues in the vicinity of those specified should also be subject to optimization of side chain conformations. In another embodiment, *Perla* can optimize all side chains. A fourth item of user input is a series of adjustable input parameters 226 that set weights for the different energy terms, place thresholds and penalties to control the flow of output and tune the effectiveness of the optimization procedure.

With this information, at step 230, *Perla* places amino acid residues from list 214 into the protein structure at positions specified in the list of residues 210. Side chains that correspond to the list of amino acids to model are obtained from a rotamer library 232. This structure is an initial form of the mutated structure and it can be validated, at step 234, by calculating various terms in the solution score.

As described in Figure 3 the intrinsic energy term of the scoring function, i.e., the intrinsic energy of interaction between the rotamers and the template protein structure, is computed, for validated structure 300 step 302. The intrinsic energy is determined by summing and optimizing van der Waals, electrostatic, and hydrogen bonding energies, computed utilizing molecular mechanics force field parameters 308 that may be read from a separate file or specified by the user. Those skilled in the art will recognize that, although the set of parameters used by the method of the present invention relate to amino acids, corresponding parameters for nucleic acids and other organic compounds, *e.g.*, carbohydrates, are available and can readily be integrated into the method as applied to other types of macromolecules.

Main chain entropy costs, vibrational side chain entropies, intrinsic contribution to the solvation energies and a penalty function are added to the molecular mechanics terms. Intrinsic energies for both the mutated structure 312 and a reference structure 306 are computed. In a preferred embodiment, the intrinsic energies for each side-chain and all of the contributions to the intrinsic energy, including, for example the solvation energy for each side-chain, are stored in memory.

In step 314, one or more selection criteria are applied to side chain conformers and those that are not compatible with the template structure are abandoned 310. In a preferred embodiment, a first criterion is an energy threshold, above which rotamers are rejected and a second criterion is a weighting from the partition coefficient, below which rotamers are rejected.

Subsequently, as described in Figure 4, at step 402 pairs of selected rotamers 400 that were not rejected at step 314 are considered in order to evaluate the pairwise (side chain-side chain) component of the scoring function. As for the intrinsic energy computation, molecular mechanics force field parameters 410 are utilized in order to compute van der Waals, electrostatic, and hydrogen bonding energies. Molecular mechanics terms are summed and optimized, step 404, and then a pairwise solvation contribution and vibrational entropy and statistical penalty is added for both mutated structure 412 and reference structure 406. No elimination need be performed at this step,

since the identification of an energetically disfavored pair does not necessarily imply that the participating side chains are incompatible with the target protein fold. In a preferred embodiment, the pairwise energies for each pair of sidechains and all of the contributions to the pairwise energy for each pair of side chains, including the solvation energy contribution for each side-chain, are stored in memory. Although this can be a very large matrix, it is better to try to store it than to repeatedly recompute the pairwise energies. For a very large contribution, for example comprising many variable side chains and utilizing a genetic algorithm, an upper limit threshold is placed on the number that are stored. In one embodiment, this may be 1 Gigabyte of random access memory. In a preferred embodiment, when a multi-processor machine is available to run *Perla*, calculation of the pairwise energies can be carried out in parallel.

In order to reduce the number of sequences to sample and to ultimately find that which has the optimal sequence-to-structure relationship, *e.g.*, the lowest native state potential energy (lowest score) or the greatest energy difference in energy from the reference state, dead-end elimination is used to establish which sequences cannot achieve the energy minimum, step 408. Those sequences that cannot achieve a desired energy minimum are marked and abandoned, 414. In an alternative embodiment, step 408 comprises the selection of a subset of sequences instead of a rejection and is achieved with the use of a genetic algorithm which gives a predefined number of best sequences. In a preferred embodiment, when a multi-processor machine is available, the dead end elimination can be run in parallel.

In one embodiment, the intrinsic and pairwise energies can be added to one another to give an estimate of the stability of the structure. This is mainly of interest for small macromolecules or where only a small number, say up to 3 residues, are being mutated and it is not worth the overhead to go through a dead end elimination, genetic algorithm or mean field theory calculation.

For all remaining sequences, mean field theory, step 416, enables the estimation of weights for all side chain rotamers, which are then used to compute the solution score 418 of each sequence. Sequences that do not score well are rejected 420, while for others, the solvation term is re-evaluated 422. Some sequences may also be eliminated at this step if they have poor solvation energies.

Finally, in the output 424 from the program, the solution structures of the engineered sequences are accompanied by a description of the energy terms that contribute to the scoring

function and a set of three-dimensional Cartesian coordinates that describe the modeled solution structure.

In two further embodiments, shown in Figure 5, a genetic algorithm is utilized. In both embodiments, the pairwise energies are not pre-computed, but are built up 512 through the use of the genetic algorithm. The selected rotamers 400 are passed directly into the genetic algorithm, 510, which simultaneously optimizes the pairwise energy 502 and the sequences. The usual molecular mechanics force field parameters 504 are used. Those contributions to the pairwise energy that have not been computed when the Mean Field Theory calculation 516 is to be carried out, are then computed 514, so that a complete set of pairwise energies is stored. In one embodiment, one hundred sequences from the genetic algorithm for example, are subjected to Mean Field Theory. The difference between the two further embodiments shown in Figure 5 that utilize a genetic algorithm is principally in the way in which the solvation energy is treated. In one embodiment in Figure 5, the intrinsic and pairwise contributions to the solvation energy are utilized within the genetic algorithm and subsequently in the Mean Field Theory calculation. Once a set of sequences has been derived, the solution scores 518 are optionally presented. In a preferred embodiment, the intrinsic and pairwise contributions to the solvation energy are subtracted out and the final solution scores computed with refined solvation energies. In an alternative embodiment, the intrinsic and pairwise energy contributions to the solvation energy are not utilized in the genetic algorithm calculation. Instead, the refined values of the solvation energy are utilized throughout. In a preferred embodiment, when a multi-processor machine is available to run *Perla*, the genetic algorithm can be run in parallel.

In yet another embodiment, a genetic algorithm may be applied to those sequences that have been obtained through the use of a dead end elimination step. In a further embodiment, it is possible to utilize a genetic algorithm to derive families of sequences without separating out the intrinsic and pairwise interaction energies. In such an embodiment, a genetic algorithm utilizes single composite solution score, to evaluate the molecular mechanics interactions, the solvation energies, and the statistical penalty.

5.3. THE GENERAL FORM OF THE SCORING FUNCTION

Central to the operation of *Perla* is the use of an scoring function to calculate the energy difference between a mutated version of the target protein and some corresponding reference

state. The way in which the reference state is constructed for proteins is described in more detail below, section 5.7.

The contributions are regarded to be either components of a molecular mechanics model or part of an empirical description of solvation and entropic factors. The theory behind each of these categories is described later in this section.

In one embodiment, the general form of the solution score consists of 6 contributions, as shown in equation (1):

$$E_{\text{Scoring Function}} = E^{\text{Molecular Mechanics}} + E^{\text{Entropy}}_{\text{Main Chain}} + E^{\text{Vibrational Entropy}}_{\text{Side Chain}} + E^{\text{Entropy}}_{\text{Side Chain}} + E^{\text{Solvation}} + E^{\text{Statistical Penalty}}_{\text{Residues}} \quad (1)$$

Each of the components in equation (1) includes as a coefficient, a user-defined weighting factor, w, which can be adjusted to suit different applications. The reason for this is that some of the terms overlap with one another in the contributions that they model, and therefore represent over-estimates of the contributions. In a preferred embodiment, w for solvation is 1.0 and all other w's are set to 0.5. Depending upon the application of the program, the coefficients can be adjusted to achieve a desired result. The explicit form of the terms is as follows.

The molecular mechanics term describes long-range interactions between pairs of atoms and supplies the difference in such terms between the target protein and the reference structure. In one embodiment, the molecular mechanics terms comprise three types of contribution, van der Waals, electrostatics and hydrogen bonding terms, each of which is multiplied by a separate coefficient w^{vdw}, w^{hb} and w^{elec}, respectively, see equations (2a,b,c). These coefficients may be adjusted to permit force field parameters from different sources to be used. For example, some published force-fields do not have separate hydrogen bonding terms. If parameters from such force fields are used, w^{hb} can be set to 0.

$$w^{vdw}\left(\sum_{\text{non-bonded atoms } i,j}\left(\frac{A_{ij}^{vdw}}{r_{ij}^{12}}-\frac{B_{ij}^{vdw}}{r_{ij}^{6}}\right)_{\text{target structure}}-\sum_{\text{non-bonded atoms } i,j}\left(\frac{A_{ij}^{vdw}}{r_{ij}^{12}}-\frac{B_{ij}^{vdw}}{r_{ij}^{6}}\right)_{\text{reference structure}}\right) \quad (2a)$$

$$+w^{hb}\left(\sum_{\text{H-bonded atoms } H,A}\left(\frac{A_{HA}^{hb}}{r_{HA}^{12}}-\frac{B_{HA}^{hb}}{r_{HA}^{10}}\right)_{\text{target structure}}-\sum_{\text{H-bonded atoms } H,A}\left(\frac{A_{HA}^{hb}}{r_{HA}^{12}}-\frac{B_{HA}^{hb}}{r_{HA}^{10}}\right)_{\text{reference structure}}\right) \quad (2b)$$

$$+w^{elec}\left(\sum_{\text{atoms, } i,j}\left(\frac{q_i q_j e^2}{4\pi\,\varepsilon_0\,\varepsilon_r\,r_{ij}}\right)_{\text{target structure}}-\sum_{\text{atoms, } i,j}\left(\frac{q_i q_j e^2}{4\pi\varepsilon_0\,\varepsilon_r\,r_{ij}}\right)_{\text{reference structure}}\right) \quad (2c)$$

In practice the molecular mechanics terms in equations (2a,b,c) are averaged over all the rotamers of each amino acid residue.

The second term in equation (1) describes the entropy cost of fixing the main chain at a physical temperature, T_{phys}, as shown in equation (3). The temperature is typically in the range 278-328K, and in a preferred embodiment is 298K. This term is akin to a secondary structure propensity term and is computed as part of the intrinsic energy contribution. For each

residue, i, in expression (3), the contributions represent effective averages over rotamers because they depend only on the identity of the amino acid.

$$-w_{\text{mainchain}}^{\text{entropy}} RT_{\text{phys}} \sum_{\text{all residues } i} \ln \frac{\sum_{\substack{\text{subspaces } \phi\psi_{20°\times 20°} \\ \text{close to } \phi_i\psi_i}} w_{\phi\psi_{20°\times 20°}} N_{\phi\psi_{20°\times 20°}}^{\text{amino acid } i}}{N_{\text{all}\,\phi\psi}^{\text{amino acid } i}} \quad (3)$$

The form of expression (3) is chosen so that it resembles an effective free energy term, ΔG, given by -RT lnK. The third term in equation (1) represents the entropy cost of restricting the "vibrational" freedom of rotamers and is shown in equation (4). It allows priority to be given to side chain rotamers that can freely rotate within a space corresponding to the Gaussian distributions determined during the creation of the rotamer library. The w_s are obtained from a partition function over the sub-rotamers of the rotamers. The w_r are obtained from Mean Field Theory. The vibrational entropy term is calculated according to expression (4).

$$-w_{\text{side chain}}^{\text{vibration}} T_{\text{phys}} \sum_{\text{all residues } i} \sum_{\substack{\text{all rotamers } r \\ \text{of residue } i}} w_r \left(\left(-R \sum_{\substack{\text{all sub-rotamers } s \\ \text{of rotamer } r}} w_s \ln w_s \right)_{\text{target structure}} - \left(-R \sum_{\substack{\text{all sub-rotamers } s \\ \text{of rotamer } r}} w_s \ln w_s \right)_{\text{reference structure}} \right) \quad (4)$$

The fourth term in equation (1) represents the entropy cost of placing amino acid side chains into the template structure where they are more hindered due to the compact protein environment. This term, expressed in equation (5), is again a difference between the entropies of the side chains in the template and the reference. The values of the w_r are obtained from Mean Field Theory.

$$-w_{\text{sidechain}}^{\text{entropy}} T_{\text{phys}} \sum_{\text{all residues } i} \left(\left(-R \sum_{\substack{\text{all rotamers } r \\ \text{of residue } i}} w_r \ln w_r \right)_{\substack{\text{target} \\ \text{structure}}} - \left(-R \sum_{\substack{\text{all rotamers } r \\ \text{of residue } i}} w_r \ln w_r \right)_{\substack{\text{reference} \\ \text{structure}}} \right) \quad (5)$$

The fifth term in equation (1) is the solvation energy. The solvation energy is computed as a difference in the energy of interaction between the target structure and surrounding solvent and the reference structure and surrounding solvent, as shown in equation (6).

$$+w^{\text{solvation}} \left(\left(\sum_{\text{atoms } i} \sigma_i ASA_i \right)_{\substack{\text{target} \\ \text{structure}}} - \left(\sum_{\text{atoms } i} \sigma_i ASA_i \right)_{\substack{\text{reference} \\ \text{structure}}} \right) \quad (6)$$

In a preferred embodiment, the solvation energy is computed over only one conformation of the solution structure, typically that obtained by using the most favorable rotamer of each mutated residue.

The last term in equation (1) is a statistical penalty function which is related to the identity of the amino acid residues in the sequence, as shown in equation (7). This term is introduced to drive the sequence design towards a sequence subspace known *a priori* to be plausible. It is not computed with respect to a reference structure of the mutated protein.

$$-w^{\text{stat}} RT_{\text{stat}} \sum_{\text{all residues } i} \ln P_{\text{amino acid } i}^{\text{stat}} \quad (7)$$

We note that the entropy of the side chains is dependent upon the weight distribution calculated by the mean field approximation routine (section 5.10). Hence, that part of the energy is *not* included at all in either the intrinsic or pairwise description. By contrast, the vibration entropy cost is used to penalize rotamers whose interaction energy (either intrinsic

or pairwise) is only optimal for a few of the sub-rotamer conformations they can adopt (see section 5.8).

For example, the values, $P^{stat}_{\text{amino acid } i}$, of this term can represent amino acid relative abundance in the protein database (PDB) or can be obtained from sequence alignments related to the family of proteins containing the target protein. The effective temperature, T_{stat}, associated with this term, might differ from the actual physical temperature T_{phys} used for entropy related terms. The factor RT_{stat} has been introduced to equation (7) to ensure that the penalty term as a whole has dimensions of energy. In another embodiment, equation (7) is extended to include terms that describe the relative occurrence of pairs of amino acids. For example, during sequence alignment, it may be found that 10% of the sequences have alanine and valine at a pair of positions, but the natural abundance of alanine and valine, respectively, may be each less than 10%. An appropriate factor, P, for a pair of residues would be the abundance of the pair (in this case 0.10) divided by the product of their individual abundances. In this way, where two residues are especially favored at a pair of positions, the value of P is greater than 1.0 and its log is a positive number. Accordingly the statistical penalty will contribute negatively, i.e., favorably to the energy. Where the value of P is 1.0, i.e., the pair of residues is neither favored/nor disfavored, the penalty does not contribute at all because the log of 1.0 is zero. Pairs of residues found to be a fixed distance apart in a sequence are often found to be close in contact in the folded structure. In yet another embodiment, equation (7) can be further extended to include to include terms describing the relative occurrence of triplets, quadruplets, etc., of amino acids.

In another embodiment, for optimization with a genetic algorithm, an additional statistical penalty may be used, equation (8).

$$\sum_{I=1}^{20} K_i\left(N_i - N_i^{PDB}\right)^2 \qquad (8)$$

Where N^{PDB} is a probability of occurrence in the PDB, and N is the probability of occurrence in the sequence of the mutated sequence. The constants K, are user-defined weights, which are adjusted to give a contribution in the range of $1Kcal/mol^{-1}$.

5.4 MODIFIED FORM OF THE SCORING FUNCTION

Additionally, for the purpose of selecting, rejecting as optimizing conformers, various components of the scoring function are partitioned into "template", "intrinsic" and "pairwise" terms.

The context in which the scoring function should be viewed is that the protein comprises a fixed backbone (or backbones) of amino acid residues, some specified subset of which are to be varied. The backbone and the constant residues (including their side chains) together form the template. The side chains of the variable residues interact with the template, giving rise to the "intrinsic" energy term and amongst themselves, giving rise to the "pairwise" energy term. In the preferred embodiment of this invention, the scoring function is therefore decomposed into a sum of terms, described respectively as "template", "intrinsic" and "pairwise" equation (9). Each of these terms partitions into summed contributions equation (10).

$$E_{\text{sequence-to-structure}} = E_{\text{template}} + E_{\text{intrinsic}} + E_{\text{pairwise}} \qquad (9)$$

$$E_{\text{sequence-to-structure}} = E_{\text{template}} + \sum_{\text{all residues, } i} E^{i}_{\text{intrinsic}} + \sum_{\text{all residues, } i>j} E^{ij}_{\text{pairwise}} \qquad (10)$$

In equations (9) and (10), $E_{\text{sequence-to-structure}}$ is an energy that is computed during the steps of optimization, prior to evaluation of the solution score for the mutated structure.

In a preferred embodiment, amino acid residues are classified as "core" or "non-core" prior to computing the scoring function. This classification is relevant for computation of the electrostatic interaction energy, through use of a variable dielectric constant, and in the calculation of the solvation energy.

In other embodiments of the present invention, the scoring function may be partitioned into terms corresponding to interactions between atoms on the macromolecule and atoms on some binding partner of interest. Such terms may be printed in the output.

5.4.2. THE TEMPLATE TERM

In a preferred embodiment, in which all atoms of the template are fixed, only the molecular mechanics, main chain entropy and statistical penalty terms are computed, as shown in equation (11).

$$E_{\text{Template}} = E_{\text{Template}}^{\text{Molecular Mechanics}} + E_{\text{Template}}^{\text{Main Chain Entropy}} = E_{\text{Template}}^{\text{StatisticalPenalty}} \qquad (11)$$

In equation (11), the subscript "Template" means all atoms of the template. In an embodiment in which all atoms of the template are fixed, the entropy contributions for the side-chains of the residues are not determined. The template term does not find application in sequence prediction where the sequence of the template is the same for all mutated structures.

The template term is useful in structure validation and because when added to the intrinsic and pairwise terms, gives an Energy Contribution that is proportional to the size of the protein being mutated.

5.4.2. THE INTRINSIC TERM

The invention provides an intrinsic energy term for all candidate rotamers, that represents the interaction of each with the main chain (and any other side chain that is kept fixed). The intrinsic energy can be used to reject unfavorable side chain rotamers.

The intrinsic energy term consists of 5 contributions, each pertaining to interactions of the side chains of the variable residues with the template.

$$E_{\text{Intrinsic}} = E_{\text{Side Chain - Main Chain}}^{\text{Molecular Mechanics}} + E_{\text{Main Chain}}^{\text{Entropy}} + E_{\text{Side Chain}}^{\text{Vibrational Entropy}} + E_{\text{Side Chain}}^{\text{Solvation}} + E_{\text{Residues}}^{\text{Statistical Penalty}} \qquad (12)$$

The terms in equation (12) mirror those in the general expression for the energy, equation (1). The intrinsic energy is expressed as a summation over all mutated residues and all those that

are flexible. The molecular mechanics contribution to the intrinsic energy for a candidate rotamer of one residue is in equations (13a,b,c), as follows.

$$w^{vdw}_{Intrinsic}\left(\sum_{\substack{\text{non-bonded}\\ \text{atoms } i \text{ of side chain}\\ \text{and } j \text{ of main chain}}}\left(\frac{A^{vdw}_{ij}}{r^{12}_{ij}}-\frac{B^{vdw}_{ij}}{r^{6}_{ij}}\right)_{\substack{\text{target}\\ \text{structure}}} - VDW^{\substack{\text{residue}\\ \text{type}\\ \text{reference}\\ \text{structure}}}\right) \quad (13a)$$

$$+w^{hb}_{Intrinsic}\left(\sum_{\substack{\text{H-bonded}\\ \text{atoms H or A of side chain}\\ \text{and H or A of main chain}}}\left(\frac{A^{hb}_{HA}}{r^{12}_{HA}}-\frac{B^{hb}_{HA}}{r^{10}_{HA}}\right)_{\substack{\text{target}\\ \text{structure}}} - HB^{\substack{\text{residue}\\ \text{type}\\ \text{reference}\\ \text{structure}}}\right) \quad (13b)$$

$$+w^{ele}_{Intrinsic}\left(\sum_{\substack{\text{atomic}\\ \text{atoms } i \text{ of side chain}\\ \text{and } j \text{ of main chain}}}\left(\frac{q_i q_j e^2}{4\pi\varepsilon_o\varepsilon_r r_{ij}}\right)_{\substack{\text{target}\\ \text{structure}}} - ELE^{\substack{\text{residue}\\ \text{type}\\ \text{reference}\\ \text{structure}}}\right) \quad (13c)$$

The molecular mechanics terms for a given rotamer may derive from averaging over all the sub-rotamers of that rotamer.

The portion of the molecular mechanics energy measured in the reference structure i.e., *VDW*, *HB* and *ELE* is dependent only on the amino acid type, not its geometry, and thus is the same for each rotamer of a given residue. The role of the reference term is to help determine which sequences might be poor quality and not to distinguish between rotamer combinations of a particular sequence. The values of the reference energies in equations (13a, b, c) that are used in the molecular mechanics term are derived from calculations done with *Perla* over a large sample of main chain structures and sequences, whose results are averaged.

The second term in equation (12), the main chain entropy cost, is also completely independent of the rotamer configuration and thus is an aid to distinguishing between sequences only. This term is computed just once for each sequence and is conveniently expressed as part of the intrinsic energy. The expression for the contribution to the main chain entropy from a given amino acid residue is in equation (14).

$$-w_{\text{mainchain}}^{\text{entropy}} RT_{phys} \ln \frac{\sum_{\substack{\text{subspaces } \phi\psi_{20^\circ x 20^\circ} \\ \text{close to residue } \phi\psi}} w_{\phi\psi_{20^\circ x 20^\circ}} N_{\phi\psi_{20^\circ x 20^\circ}}^{\substack{\text{residue} \\ \text{type}}}}{N_{\text{all}\,\phi\psi}^{\substack{\text{residue} \\ \text{type}}}} \qquad (14)$$

The third term in equation (12), describes the side chain rotamer vibration entropy cost is measured with respect to a set of tabulated references. The weights of the sub-rotamers are obtained from the partition function. The contribution made by one rotamer to the side chain vibrational entropy component of the intrinsic energy, is expressed in equation (15).

$$-w_{\text{side chain}}^{\text{vibration}} T_{\text{phys}} \left(\left(-R \sum_{\text{sub-rotamers } s} w_s \ln w_s \right)_{\substack{\text{target} \\ \text{structure}}} - VIB_{\substack{\text{reference} \\ \text{structure}}}^{\substack{\text{residue} \\ \text{type}}} \right) \qquad (15)$$

In a preferred embodiment the vibrational contribution to the reference structure is calculated from a uniform distribution. In equation (16), N is the number of sub-rotamers, which can be adjusted by the user.

$$VIB_{\substack{\text{reference} \\ \text{structure}}}^{\substack{\text{residue} \\ \text{type}}} = -RT \sum_{\text{sub-rotamers}} N_{\text{sub-rotamers}}^{-1} \ln N_{\text{sub-rotamers}}^{-1} \qquad (16)$$

In another embodiment, the reference term in equation (15) can be computed from a number of calculations of *Perla* on a database of peptide fragments or, using a Gaussian Distribution.

The fourth term in equation (12) which measures the solvation energy has been obtained by cutting the surface areas into intrinsic and pairwise parts. The contribution that one residue makes to the solvation component of the intrinsic energy is given in equation (17).

$$+w_{\text{intrinsic}}^{\text{solvation}} \sum_{\text{atoms } i \text{ of side chain}} \sigma_i \left(\left(ASA_i \right)_{\substack{\text{reference} \\ \text{structure}}} - \left(ASA_i \right)_{\substack{\text{target} \\ \text{structure}}} \right) \qquad (17)$$

The solvation term is expressed from the relative buried surface area rather than the exposed surface area (thus the invention provides a subtraction in the sense of *reference-target* and not *target-reference*). For this reason, a different set of solvation parameters σ_i is used from those used to compute the solvation contribution to the scoring function energy of equation (1), as described later.

Finally, the fifth term in equation (12) is a statistical contribution, which should consist of tabulated values, propensities or probabilities given by the user in a readable format. The contribution to this term in the intrinsic energy made by one residue is given in equation (18).

$$-w^{stat}_{Intrinsic} RT_{stat} \ln P^{stat}_{\substack{residue \\ type}} \qquad (18)$$

5.4.3. THE PAIRWISE TERM

Finally, the pairwise energy term represents the interaction energy of a pair of candidate rotamers and is therefore summed over all pairs of candidate rotamers. The previous comments about the reference states and temperatures are applicable here.

The pairwise term comprises 4 contributions, equation (19):

$$E_{Pairwise} = E^{Molecular\ Mechanics}_{Side\ Chain-Side\ Chain} + E^{Vibrational\ Entropy}_{Side\ Chain} + E^{Solvation}_{Side\ Chain} + E^{Statistical\ Penalty}_{Residues} \qquad (19)$$

Similarly to the Intrinsic term, the molecular mechanics contribution to the pairwise term is split up into weighted contributions from van der Waals, electrostatic and hydrogen bonding energies. The summations run over pairs of atoms on different residue side chains, equations (20a,b,c).

$$w^{vdw}_{Pairwise}\left(\sum_{\substack{\text{non-bonded}\\ \text{atoms } i \text{ of first side chain}\\ \text{and } j \text{ of second side chain}}}\left(\frac{A^{vdw}_{ij}}{r^{12}_{ij}}-\frac{B^{vdw}_{ij}}{r^{6}_{ij}}\right)_{\substack{\text{target}\\ \text{structure}}} - VDW^{\substack{\text{residue}\\ \text{pair}}}_{\substack{\text{reference}\\ \text{structure}}}\right) \quad (20a)$$

$$+w^{hb}_{Pairwise}\left(\sum_{\substack{H\text{-bonded}\\ \text{atoms } H \text{ or } A \text{ of first side chain}\\ \text{and } H \text{ or } A \text{ of second side chain}}}\left(\frac{A^{hb}_{HA}}{r^{12}_{HA}}-\frac{B^{hb}_{HA}}{r^{10}_{HA}}\right)_{\substack{\text{target}\\ \text{structure}}} - HB^{\substack{\text{residue}\\ \text{pair}}}_{\substack{\text{reference}\\ \text{structure}}}\right) \quad (20b)$$

$$+w^{ele}_{Pairwise}\left(\sum_{\substack{\text{atomic}\\ \text{atoms } i \text{ of first side chain}\\ \text{and } j \text{ of second side chain}}}\left(\frac{q_i q_j e^2}{4\pi\varepsilon_o\varepsilon_r r_{ij}}\right)_{\substack{\text{target}\\ \text{structure}}} - ELE^{\substack{\text{residue}\\ \text{pair}}}_{\substack{\text{reference}\\ \text{structure}}}\right) \quad (20c)$$

The molecular mechanics contribution to the pairwise energy also contains reference structure terms, i.e., *VDW*, *ELE* and *HB*, which depend only on amino acid types and are obtained by averaging over a number of different conformations obtained from a database.

The second term in equation (19), for the rotamer vibration entropy, is formulated to measure the change of entropy due to the interaction of the two side chain rotamers taking as a reference the vibration entropy of each rotamer substituted separately in the target structure, as shown in equation (21).

$$-w_{\text{Pairwise}}^{\text{vibration}} T_{\text{phys}} \lambda \left(\left(-R \sum_{\substack{\text{sub-rotamers } A_s \text{ and } B_s \\ \text{of rotamer pair } AB}} w_{A_s B_s} \ln w_{A_s B_s} \right)_{\substack{\text{rotamer pair } AB \\ \text{in target structure}}} - \left(\left(-R \sum_{\substack{\text{sub-rotamers } A_s \\ \text{of rotamer } A}} w_{A_s} \ln w_{A_s} \right)_{\substack{\text{only rotamer } A \\ \text{in target structure}}} + \left(-R \sum_{\substack{\text{sub-rotamers } B_s \\ \text{of rotamer } B}} w_{B_s} \ln w_{B_s} \right)_{\substack{\text{only rotamer } B \\ \text{in target structure}}} \right) \right) \quad (21)$$

This entropy term is scaled by a factor λ to avoid an overestimation when summing over all pairs of interacting rotamers (see section 5.8 for details).

The third term in equation (19), for the difference between accessible surface areas is formulated to measure the area buried between the two side chains. This solvation term, equation (22), is now also scaled by a factor λ_s to avoid an overestimation of the buried surface areas (likely to be counted several times when summing over all pairs of interacting rotamers).

$$w_{\text{pairwise}}^{\text{solvation}} \sum_{\substack{\text{atoms } i \text{ of} \\ \text{residue } A \text{ or } B \text{ of} \\ \text{residue pair } AB}} \sigma_i \lambda_s \left((ASA_i)_{\substack{\text{only residue } A \text{ or } B \\ \text{in target structure}}} - (ASA_i)_{\substack{\text{Residue pair } AB \\ \text{in target structure}}} \right) \quad (22)$$

The final term in equation (19) is, as previously, introduced to bias against improbable sequences of residues, though comprises a term to represent the probability of a pair of residues being present in the structure. The contribution for each pair of residues is expressed in equation (23).

$$-w^{stat}_{Pairwise} RT_{stat} \ln P^{stat}_{\substack{residue \\ pair}} \qquad (23)$$

5.5. THE MOLECULAR MECHANICS POTENTIAL

In the method of the present invention, a protein is represented as an ensemble of atoms with discrete masses and partial charges and, therefore, classical mechanics equations are applied to estimate the potential energy of the system.

The standard molecular mechanics function (or "force field") is a sum of terms that are related to bonded or nonbonded interactions and that depend on the atomic configuration, which is described by the coordinate vectors, r_i (for an overview, see van Gunsteren & Berendsen, 1990, *Angew. Chem. Int., Ed. Engl.* 29:992-1023), in equations (24a,b,c,d):

$$V(r_1,\ldots r_n) = \Sigma_{bonds} \frac{1}{2} K_b (b - b_0)^2 + \Sigma_{angles} \frac{1}{2} K_\Phi (\Phi - \Phi_0)^2 \qquad (24a)$$

$$+ \Sigma_{dihedrals} K_\Phi (1 + \cos(n\varphi - \delta)) \qquad (24b)$$

$$+ \Sigma_{nonbonded\,ij} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \Sigma_{H-bonds\,ij} \left(\frac{A'_{ij}}{r_{ij}^{12}} - \frac{B'_{ij}}{r_{ij}^{10}} \right) \qquad (24c)$$

$$\Sigma_{charges\,ij} \frac{q_i q_j e^2}{4\pi \varepsilon_0 \varepsilon_r r_{ij}} \qquad (24d)$$

Molecular mechanics is a well-established sphere of research and several widely used implementations exist: for example, AMBER, CHARMM, ECEPP2, MM2, CVFF, all of which are commercially or freely available. The operation of *Perla* is not dependent on the specific force-field which is used, because *Perla* is only concerned with ranking sequences. In an alternate embodiment, the user is free to adjust values of molecular mechanics parameters, where, for particular atom types, the parameters from a published force field are, in some way inadequate.

5.5.1. THE BOND STRETCH AND BOND-ANGLE TERMS

The first three terms of the molecular mechanics force field correspond to bonded interactions. The first represents the elongation of covalent bonds between two atoms (bond stretching). It has a harmonic form, where b is the effective bond length, b_0 is the ideal length (energy minimum), and K_b is the force constant that is characteristic of the actual type of covalent bond. The second term similarly describes the deformation of the angle ϕ formed by three covalently bonded atoms (bond-angle bending). The third accounts for the rotation around bonds, or dihedral angles φ, according to a periodic potential with phase δ.

The description of side chain conformations as a set of rotamers consists of setting the corresponding dihedral angles at values corresponding to energy minima. In addition, the covalent bond lengths and angles are set to their ideal values and are invariant. Therefore, in a preferred embodiment, the related energy terms are neglected and the methods of the present invention only consider the remaining three terms: van der Waals, hydrogen bonding and electrostatic interactions. These noncovalent forces that maintain protein three-dimensional structures, are the most important for a valid representation of protein structure, and are described in mathematical detail below. In a preferred embodiment, all parameters, *e.g.*, atomic charges and van der Waals energy parameters, are taken from the ECEPP/2 potential (Momani *et al.*, 1975, *J. Phys. Chem.* 79:2361-2381; Nemethy *et al.*, 1983, *J. Phys. Chem.* 87:1883-1887). Although molecular mechanics force fields have been parameterized with goals other than protein design in mind, their application to protein design in the methods of the present invention can be justified because the overall scoring function contains other terms, such as entropic, solvation and statistical penalty terms, to account for respectively entropy, solvation and statistical bias, and because the contribution of the various terms can be adjusted by user-defined weighting factors.

5.5.2. THE VAN DER WAALS ENERGY

Van der Waals interactions originate from a nonspecific attractive force that exists between atoms. That force is due to the transient asymmetry of the distribution of electronic charge around an atom, which induces a similar asymmetry in the distribution of electronic charge around neighboring atoms. The attraction increases as the distance between atoms decreases, until it is at a maximum when the two atoms, *i* and *j*, are separated by a distance r_{ij}, which is about 0.3-0.5 Å larger than the van der Waals contact distance, (the closest contact distance between the two atoms that is observed in crystal structures). The overlapping of the electron clouds of atoms *i* and *j* creates strong dominant repulsions at shorter distances (Fig. 6).

The van der Waals interaction energy between two atoms *i* and *j* can be described by a standard 6-12 Lennard-Jones potential, and the total van der Waals interaction energy term is the sum of the interaction energies between all nonbonded atom pairs:

$$E_{vdw} = \sum_{\text{nonbonded}\, ij} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \qquad (25)$$

where r_{ij} is the distance separating atoms *i* and *j*, and A_{ij} and B_{ij} are related to the type and environment of the interacting atoms. Thus, the energy term consists of a repulsive part that decays with r^{12}, and an attractive part that varies inversely with r^6.

Van der Waals energies for pairs of atoms are on the order of the average thermal energy of molecules at room temperature (~0.6 kcal/mol) and diminish rapidly even for a small increase of interatomic distances. Thus, the van der Waals interaction becomes significant only when many interacting pairs form simultaneously, such as in the folded state of a protein. Most importantly, van der Waals energies are critical probes of the packing quality within the three-dimensional fold. For any sequence to fit a given fold, steric compatibility is required and no positive van der Waals energies should be tolerated. Cavities, which produce van der Waals energies near zero, should also be avoided, especially if they are small enough to exclude solvent water molecules.

In the reference (denatured) state, atomic contacts within the polypeptide are less common and intra-molecular van der Waals interactions are not as significant as in the folded state, since van der Waals contacts with the solvent partly compensate for the loss of interaction. Therefore, most existing computer-based methods for designing proteins neglect the van der Waals contribution of the denatured state.

However, consideration of the van der Waals energy of the reference state of a protein in the method of the present invention results in the removal of scaling artefacts in the van der Waals energy term. Potential energy functions that sum over all atoms of a system scale with the number of interacting atoms, resulting in energy terms that contain large numbers. These scaling artefacts that should be avoided. In the method of the present invention, scaling artefacts are avoided when comparing two sequences with different amino acid compositions by referencing each sequence to a denatured conformation. In the preferred embodiment of the present invention, van der Waals reference energies for each amino acid are utilized. These may be obtained in several different ways. In one approach, energy terms were calculated for each of the twenty amino acids in an extended five-residue peptide with alanine residues flanking the residue of interest. In another approach, energy terms are calculated for each of the amino acids, as found in a population of protein fragments similar to the population of unfolded structures. The reference energies scale with the number of atoms in each amino acid and compensate for the larger van der Waals contribution of larger residues in folded proteins.

5.5.3. THE ELECTROSTATIC ENERGY

The interaction energy between electrostatic charges is fundamental and is described as the sum over all nonbonded atoms *i* and *j*, as follows, equation (26)

$$E_{\text{elec}} = \sum_{\text{charges } ij} \frac{q_i q_j e^2}{4\pi \varepsilon_0 \varepsilon_r r_{ij}} \quad (26)$$

wherein q_i and q_j are the numbers of charges on atoms *i* and *j*, respectively, r_{ij} is the distance between the two atoms, *e* is the charge of an electron, and ε_0 and ε_r are the permittivity of the vacuum and the medium relative dielectric constant, respectively.

In a vacuum, the electrostatic potential of an atomic charge in the field of another is the product of the two atomic charges divided by the distance that separates them (from Coulomb's Law). For two charges of opposite sign, the energy decreases as the atoms approach each other; the energy increases with decreasing distance if the charges have the same sign. The interaction is strong, *e.g.*, up to 100 kcal/mol, and is effective over large distances.

In media other than vacuum, the strength of the interaction is significantly reduced to less than several kcal/mol by the relative dielectric constant ε_r. In one embodiment, the ε_r of

water has a value of about 80; the interior of a protein, which is mainly packed with carbon atoms, is less polar and has a lower dielectric constant, usually 4.0.

In the method of the present invention, two dielectric constants (one for water or bulk solvent and one for the interior of the protein) are used for each mutated residue, according to the degree of burial of the side chain at the related position in the target protein structure. Every side chain atom is determined to be "non-core", i.e., exposed or "core", i.e., buried according to some geometric criterion. In one embodiment, this criterion is derived from the relative proportion of the solvent accessible surface area of the side chain of the input structure that is assigned to the atom. In a preferred embodiment, the geometric description takes into account the distance from Cα to the nearest solvent molecule on a solvent accessible surface constructed with an 8Å probe, taken along the Cα - Cβ vector, as well as the shortest distance from the Cα atom to any solvent molecule on that surface. For each pair of atoms for which an electrostatic interaction is being calculated, the dielectric constant used will be the solvent value if both atoms are exposed, the protein interior value if both atoms are buried. When the interaction is between a completely buried atom and a completely exposed atom, the average of both dielectric constants is used.

In addition, the electrostatic energy term is modified to lessen the importance of the interaction at long distance. In one embodiment, shown in Equation 27, dielectric constants are scaled linearly with the separation distance between atoms *i* and *j*:

$$\varepsilon_r = \varepsilon r_{ij} \qquad (27)$$

In the preferred embodiment of the present invention, the pH is considered to be neutral, and the parameters used are for fully charged versions of acidic (aspartic acid, pK=3.5; glutamic acid, pK=4.5; histidine, pK=6) and basic (lysine, pK=11; arginine, pK=12) amino acids. Therefore, the entire electrostatic energy term is scaled by an exponential factor to account for the screening of charges by salts and counterions, as shown in Equation 28:

$$E_{\text{elec}} = \sum_{\text{charges } ij} \left(\frac{q_i q_j e^2}{4\pi\varepsilon_0 r_{ij}} \right) \left(\exp \frac{-r_{ij}}{r_s} \right) \qquad (28)$$

The rate of exponential damping is controlled by a decay constant, r_s, whose units are those of distance. The decay constant, r_s, in equation (28) can be calculated according to methods

described in Lacroix E., Viguera A.R. and Serrano L. 1998 J. Mol. Biol. 284:173-191. An expression for r_s is given in Equation (29):

$$\frac{1}{r_s} = \left(\frac{8\pi e^2 N_A I}{1000kT} \right)^{\frac{1}{2}} \qquad (29)$$

In equation (29), N_A is Avogadro's number, I is the ionic strength of the solution and k is Boltzmann's constant.

When the electrostatic interaction energy is computed according to equations (26) and (27) or (28) the preferred value of ε_r is between 8.0 and 32.0. Fig. 7 illustrates the electrostatic interaction energy between two unit charges of equal sign as a function of interatomic separation calculated using either Equations (26) and (27) or equation (28).

In the denatured state, electrostatic energy terms contribute less to the potential energy due to the overall increase in interatomic distances caused by extension of the protein chain, and more importantly, to higher solvation and charge screening. In contrast, in the folded protein, amino acids separated by only a few residues rarely undergo a conformational change that gives rise to a significant change in the distances separating their atomic charges. Therefore, if only the native state of the protein is considered, the electrostatic term is over-estimated. In a preferred embodiment of the present invention, values are tabulated to represent all possible electrostatic interactions in the denatured state as a function of the sequence separation. In one embodiment of the present invention, the electrostatic energy term of the reference state is zero.

5.5.4. HYDROGEN BONDING ENERGY

A hydrogen bond is formed when two electronegative atoms, a donor and an acceptor, compete for the same hydrogen atom. As a result, the distance between the hydrogen atom of the hydrogen bond donor and the hydrogen bond acceptor is shorter than the van der Waals contact distance, although it is larger than the length of a covalent bond. The interaction is partly covalent and partly electrostatic in nature and can have an energy of up to 7 kcal/mol. Hydrogen bonds are highly directional and occur predominantly with the donor, hydrogen, and acceptor in a collinear orientation. Therefore, the potential energy

function of the preferred embodiment of the present invention considers hydrogen bonding only if the geometrical conditions are satisfied, *i.e.*, if the distance between the hydrogen and the acceptor atom is between 1.7 Å and 2.5 Å, and the angle made by the donor, hydrogen and acceptor is greater than 100° (Fig. 4). If these conditions are met, a hydrogen bonding term (Equation 30) replaces the van der Waals term corresponding to the interaction between the hydrogen and acceptor atoms.

$$E_{\mathrm{HB}} = \sum_{\text{H-bonded } H,A} \left(\frac{A_{HA}}{r_{ij}^{12}} - \frac{B_{HA}}{r_{ij}^{10}} \right) \qquad (30)$$

The preferred embodiment of the present invention does not take into account the possibility that there is intra-molecular hydrogen bonding in the denatured state, since the geometrical conditions are only fulfilled if elements of structure, *e.g.*, turns or α-helices, form locally. The formation of hydrogen bonds between atoms in the denatured protein and water is included empirically in an accessible surface-area-dependent solvation potential described below in Section 5.6. In essence, the residues in a denatured protein are modeled by ensembles of representative fragments taken from protein structures in the PDB.

5.6. THE EMPIRICAL POTENTIAL

The method of the present invention, as implemented in the preferred embodiment, *Perla*, also evaluate changes in entropy and solvation of the protein chain by means of empirical models constructed to account for properties that cannot be broken down into a set of well characterized physical forces. It is customarily difficult to accurately model entropy and solvation, because a practical and accurate representation of the ensemble of unfolded protein structures would have to be developed. This would necessitate the handling of an enormous number of either water molecules or chain configurations within a practical amount of computing time, as well as the development of an accurate set of energy parameters to describe these unfolded states. Instead *Perla* adopts pragmatic levels of approximation.

5.6.1. SOLVATION

Proteins function in aqueous media, which are poor solvents for apolar molecules because apolar molecules cannot participate in hydrogen bonding with liquid water. To satisfy their hydrogen bonding requirement, water molecules that surround a hydrophobic molecule order themselves by hydrogen bonding with each other, and consequently, lose many degrees of freedom. The reduction of exposed hydrophobic surfaces through protein folding leads to a release of ordered layers of water molecules, and consequently, the entropy of the solvent increases. This increase in the entropy of the system is the basis for the hydrophobic effect, which leads to proteins adopting compact shapes. In terms of protein design, the essential property of water is the partitioning of polar and apolar residues between the protein surface and interior, or core. As a result of the hydrophobic effect, apolar residues are preferentially, but not always, buried in the protein interior, where the aqueous solvent is excluded. Conversely, polar residues may occasionally be buried but are preponderantly found on the protein surface; charged residues are rarely buried.

Due to the large number of water molecules in the layers surrounding the protein surface, water cannot be explicitly modeled in order to consider the effect of solvent on sequence preferences. Eisenberg and McLachlan showed that the free energy of interaction of a protein with water could be represented by the sum of the interaction energies of each atom of the protein with solvent. They further proposed that the interaction strength is proportional to the accessible surface area, (ASA_i), of each atom.

In a preferred embodiment of the present invention, water is modeled implicitly (in bulk, rather than as discrete molecules) and the solvation potential energy term is calculated from the difference in accessible surface area of each atom *i* in the folded and denatured protein (Δ ASA_i) and from empirically determined solvation parameters for each atom (σ_i), as shown in Equation 31:

$$E_{solv} = \sum_{\text{all atoms } i} \sigma_i \Delta ASA_i \quad (31)$$

Accessible surface areas depend only on the atomic configuration of the protein and are calculated using the method of Lee and Richards (1971, *J. Mol. Biol.* 55:379-400) and the numerical surface calculation (NSC) routine of Eisenhaber *et al.* (1995, *J. Comp. Chem.* 16:273-284). Many other suitable surface area calculation algorithms are, however, known to one skilled in the art and could be utilized with the methods of the present invention. A water molecule with radius of 1.4 Å is the "probe" that is rolled along the van der Waals

surface of the protein atoms in order to calculate the accessible surface. The atomic radii and solvation parameters used to calculate the accessible surface areas of proteins in a preferred embodiment of the present invention are taken from Eisenberg and McLachlan (1986, *Nature* 319:199-203), as described in section 5.11.

However, correct accessible surface area measurements can only be made in the context of a full structure, and not before the optimal combination of side chain rotamers is found for the evaluated sequence. Therefore, in order to evaluate solvation contributions to the intrinsic and pairwise energies, an alternative approach is adopted.

It has been found that, when evaluating the intrinsic and pairwise solvation energies for different sequence variants and conformations during optimization, changes in ASA values for pairs of residues upon folding leads to an overestimation of the area of buried surfaces. Nevertheless, it is important to include a solvation parameter during optimization routines. Thus, in a preferred embodiment of the present invention, polar and apolar buried surfaces evaluated in a pairwise manner are scaled down in the manner proposed by Street and Mayo (1998, *Folding & Design* 3:253-258). The surface area of residue *i* buried by the template, given by BSA_i, is evaluated as the difference between the accessible surface area of the same residue placed at the center of a five residue peptide and the accessible surface area of the residue in the target protein conformation, is shown in equation (32):

$$BSA_i = ASA_i^{5\text{-peptide}} - ASA_i^{\text{target structure}} \qquad (32)$$

The surface area buried between residues *i* and *j* is evaluated as the difference between the exposed surface area of each residue separately placed in the target conformation and the exposed surface area of the pair of residues placed together in the target protein conformation. This is shown as follows:

$$BSA_{ij} = \lambda_s \left(ASA_i^{\text{target structure}} + ASA_j^{\text{target structure}} - ASA_{i \text{ and } j}^{\text{target structure}} \right) \qquad (33)$$

To compensate for the overestimation of total buried surface area, the ASA terms are scaled with a factor, λ_s, that depends on the location of residues *i* and *j*. Equation (30) differs from that given by Street and Mayo because the factor λ_s multiplies all the ASA terms. In one embodiment, λ_s is taken to be 0.40 for core residues, 0.75 for non-core residues, and 0.60 for a pair that consists of one core residue and one non-core residue. In a preferred embodiment, λ_s can be related to an alternative, pre-calculated parameter, Λ:

$$\lambda_s = \Lambda \frac{N_{\text{first rotamer}}^{\text{contacts}} + N_{\text{second rotamer}}^{\text{contacts}}}{N_{\text{first rotamer}}^{\text{contacts}} N_{\text{second rotamer}}^{\text{contacts}}} \qquad (34)$$

In a preferred embodiment, $\Lambda = 0.5$, though others values could be obtained by a fitting exercise. The solvation energy is obtained by summing equations 32 and 33 over all side chains and by multiplying the accessible surface areas by the solvation parameters 0.100 for polar buried surfaces and -0.026 for nonpolar buried surfaces.

5.6.2. ENTROPY OF THE MAIN CHAIN

The entropy change upon folding, another major component of protein stability, is calculated in a preferred embodiment of the present invention using a statistical approach. Although entropy is a unified physical concept, it is practical to divide the entropy change into parts related to either the main chain or to the side chains (see Section 5.9, below). The main chain entropy term is expressed as the cost to fix the backbone conformation into the ensemble of ϕ and ψ dihedral angles of the target structure. These costs depend on the nature of the amino acid located at each ϕ-ψ pair, and were predetermined for use in the preferred embodiment of the invention to reflect the secondary structure propensities of the twenty amino acids (Muñoz & Serrano, 1994, *Proteins* 20(4):301-311), as described below. The entropy of the main chain can therefore also be thought of as a secondary structure propensity term.

A set of 527 protein structures that share less than 35% sequence homology (PDBSELECT; Hobohm & Sander, 1994, *Protein Sci.* 3:522-524; Hobohm *et al.*, 1992, *Protein Sci.* 1:409-417) was used to obtain all main chain dihedral angles. The numbers of occurrences of each amino acid in regions of the Ramachandran (ϕ-ψ) plot sampled at fixed intervals, d_0 were determined. In a preferred embodiment, d_0 is taken to be twenty degrees.

The tendency for amino acid X to populate a particular region of Ramachandran space, *e.g.*, ϕ_i-ψ_i, is the ratio of the number of hits in the interval considered ($N_{\phi i\text{-}\psi i}$) and the total number of hits for amino acid X ($N_{all\ \phi\text{-}\psi}$):

$$P^{X}_{\phi_i-\psi_i} = \frac{N^{X}_{\phi_i-\psi_i}}{N_{all\phi-\psi}} \qquad (35)$$

For such a partitioning of dihedral angle space, it is also necessary to quantify the distance, *d*, between pairs of points (each of which represents a residue conformation):

$$d_{\phi\psi 20^\circ X 20^\circ} = \sqrt{\left(\phi_i - \phi_{20^\circ X 20^\circ}\right)^2 + \left(\psi_i - \psi_{20^\circ X 20^\circ}\right)^2} \qquad (36)$$

For 20° X 20° regions which are more than a threshold distance (in angle space), d_t from the residue ϕ_i - ψ_i dihedral angles, the occurrences are modified with a weight given by an exponential decay function of the separation distance, (to the inventors' knowledge, such an approach has not been used before in computer-based protein design methods):

$$w_{\phi\psi 20^\circ X 20^\circ} = \exp\left(-d_{\phi\psi 20^\circ X 20^\circ} / d_o\right) \qquad (37)$$

This modification allows a smooth transition of the energy function over the main chain dihedral angle space (instead of the abrupt changes that occurred previously when crossing the 20° x 20° boundaries). Thus, empirical observation is used to calculate the likelihood that a particular amino acid will have main chain dihedral angles ϕ_i and ψ_i in any given protein structure.

In the preferred embodiment, the database used is large enough to be considered as a system under thermodynamic equilibrium in which pseudo-energies or costs to displace the equilibrium toward a particular state can be calculated from the natural logarithm of the ratio in Equation 32. Entropy costs to fix the main chain dihedral angles of amino acid X in

a particular region of the Ramachandran plot, *e.g.*, ϕ_i-ψ_i, were therefore calculated as shown in equation (38) for use in a preferred embodiment of the present invention:

$$-RT_{phys}\ln\frac{\sum_{\text{subspaces } \phi\psi_{20^\circ x20^\circ} \text{ close to residue } \phi\psi} w_{\phi\psi_{20^\circ x20^\circ}} N^{\text{residue type}}_{\phi\psi_{20^\circ x20^\circ}}}{N^{\text{residue type}}_{\text{all } \phi\psi}} \qquad (38)$$

5.7. CONSIDERATION OF THE DENATURED STATE OF PROTEINS

An important consideration when modeling proteins is that of a reference state or configuration. In practice, proteins in solutions are dynamic ensembles of structures: the folded ("native") structures are in equilibrium with unfolded "denatured" configurations. The former are compact, with residues in close contact with non-neighboring residues due to the intricacies of the backbone configuration, whereas the latter are open, more chain-like. For the purposes of the present invention, the essential difference between these two extremes is that individual residues (particularly their side chains) participate in many more pairwise interactions amongst themselves in the folded state than in the unfolded. It is desired to quantify this difference at the level of individual residues, a result which is achievable as described below. A further reason to use a reference state configuration is so that the calculated solution store remains approximately proportional to the stability of the solution structure.

Whereas native protein structures are available (in the PDB) denatured structures must be obtained via simulation. In a preferred embodiment, a set of non-homologous proteins (obtained from the WHATIF database) is used to extract all protein fragments that are at least 4 and at most 20-residues long. These peptide segments may be clustered into groups according to length and structural homology, using a combination of main chain dihedral angle comparisons and internal (*i.e.*, Cα-Cα) distance comparisons. There are many clustering algorithms which may be used for this purpose, for example, Ward's, Jarvis-Patrick and assorted hierarchical methods. For each cluster, a single representative is selected (for example, from the geometrical center of the cluster). The ensemble of representatives is used as a set of main chain templates to reconstruct sequences of the type

$(Ala)_m$-X -$(Ala)_n$ and $(Ala)_l$-X-$(Ala)_m$-Y-$(Ala)_n$ where X and Y are any of the 20 natural amino acids (and the subscripts, *l*, *m*, *n*, represent segment lengths) and Ala represents the amino acid, alanine. (These sequences represent the amino acid residues of interest in an ordinary environment: alanine is the amino acid with the smallest (shortest) carbon containing side chain. It therefore contains no polar or bulky groups which might cause folding or twisting of the sequence but, unlike glycine (whose side chain is hydrogen), has sufficient bulk to prevent collapse.) *Perla* itself can be used to determine a solution score for each sequence, i.e., each peptide fragment. It is then possible to compute an average solution score that corresponds to the output of a partition function measured over the ensemble of fragments. *Perla* does so for each separate energy term, and then provides sets of values to be used as reference values for the random coil, *i.e.*, the denatured state. The references for the intrinsic parts of the van der Waals, hydrogen bonding and electrostatic energy terms, and the side chain rotamer entropies, are measured with sequences of the type $(Ala)_m$ -X-$(Ala)_l$, while the references for the pairwise parts of the van der Waals, hydrogen bonding and electrostatic energy terms, are measured with sequences of the type $(Ala)_l$- X-$(Ala)_m$ -Y- $(Ala)_n$. In a preferred embodiment of the present invention, when more than one CPU is available, reference energies may be computed in parallel with the energies of the folder structures.

For application to other categories of macromolecules, it may be appropriate to consider an alternative form of reference state. For example, when attempting to find a set of nucleotides that improves the interaction between a nucleic acid and a protein or other nucleic acid sequence, the reference may be the isolated nucleic acid in question, whereas the scoring function will quantify the extent of the interaction.

5.8. CONSTRUCTION OF THE ROTAMER LIBRARY USED IN *PERLA*

Crystallographically determined protein structures show that the side chain dihedral angles are not distributed uniformly through 360° (Janin & Wodak, 1978, *J. Mol. Biol.* 125:357-386; Ponder & Richards, 1987, *J. Mol. Biol.* 193:775-791). For example, the torsion angles around two bonded sp^3 carbons generally cluster into 'gauche +' (+60°), 'gauche -' (-60°), and 'trans' (180°) conformations (Fig. 6). In a preferred embodiment of the present invention, we wish to construct rotamer libraries in which all significant conformers are represented.

By way of example, the set of 527 protein structures that share less than 35% sequence homology (see Section 5.5; PDBSELECT; Hobohm & Sander, 1994, *Protein Sci.* 3:522-524; Hobohm *et al.*, 1992, *Protein Sci.* 1:409-417) was used to obtain all sets of side chain dihedral angles (χ_1, χ_2, χ_3, χ_4). For each amino acid other than glycine and alanine, the distribution $\upsilon(\chi)$ of dihedral angles determined from the Protein Data Bank structures was fitted to a combination of normal distributions represented by a sum of Gaussians, equation (39) as follows:

$$\upsilon(\chi) = k + \sum_{i=1}^{N} \left(\frac{1}{\sigma_i \sqrt{2\pi}} \right) \exp\left(\frac{-(\chi - \mu_i)^2}{2\sigma_i^2} \right) \qquad (39)$$

The number of Gaussian terms, N, was modified until no further reduction of the square of the difference between observed and calculated distributions was obtained. A constant term, k, was added to fit the distribution of side chain dihedral angles with poorly marked preferences, *e.g.*, χ_2 of Asn and Asp, or to represent the noise. The outputs of the fitting procedure are the centers (μ_i) of the N Gaussian peaks and their standard deviations, σ_i. In addition to the expected well-defined peaks of standard conformers, separate distributions of lower amplitudes were used to fit the data. In most cases, as illustrated for valine (light grey peaks in the top panel of Fig. 12), the additional Gaussian curves represent variations around different dihedral angle values that can be adopted by a particular amino acid in a significant number of instances.

Side chain rotamers with all combinations of the peak centers, μ_i (except for "ghost" peaks) were constructed using ideal values for the covalent bonds and angles (Mazur & Abagyan, 1989, *J. Biomol. Struct. Dyn.* 6(4):815-832; Momani *et al.*, 1975, *J. Phys. Chem.* 79:2361-2381; Nemethy *et al.*, 1983, *J. Phys. Chem.* 87:1883-1887). For dihedral angles that do not have a normal distribution, *e.g.*, χ_2 of asparagine and aspartic acid, and χ_3 of glutamine and glutamic acid, sets of values were chosen to sample the range of observed values. The constructed side chains were inspected visually and conformations with steric clashes were removed. The remaining side chain conformations form the custom-made rotamer library of 419 rotamers employed by the preferred embodiment of the present

invention, (Table 1). Side chains built by the preferred embodiment of the present invention are taken from this library. The advantage of this approach for the construction of a rotamer library is that it does not use stereochemical rules to generate the rotamers, thus allowing for the addition to the library of less abundant but relevant rotamers. Furthermore, the fitted normal distributions define the margins within which the rotamers can oscillate during evaluation of sequences.

The generation of dihedral angles by means of equation (39) does not include the fact that consecutive angles have correlated distributions. The correlation, due to the topological structures of certain amino acids, can be so strong that a particular value of χ_1 is only possible if χ_2 itself adopts a defined value (*e.g.*, the -95, 36 conformer of leucine in Fig. 12). Furthermore, the fitting procedure used to generate the dihedral angles for the rotamers used in the library cannot detect rare peaks.

More preferably, side chain conformations with correlated dihedral angles or rare dihedral angles can be included in the rotamer library by repeating the analysis above while considering the distribution of all dihedral angles of an amino acid simultaneously, since rare peaks are more resolved in a multidimensional representation. The rotamer library employed by the methods of the present invention contains only a few identified cases of rare or correlated dihedral angle values, *e.g.*, the -175, 150 and -145, -150 leucine conformers, (Fig. 12).

Table 1: Summarized Description of Rotamer Library

Amino Acid	Dihedral Angles	Number of Rotamers	Amino Acid	Dihedral Angles	Number of Rotamers
Ala	-	1	Met	χ_1, χ_2, χ_3	27
Cys	χ_1, χ_2	18	Asn	χ_1, χ_2	18
Asp	χ_1, χ_2	9	Pro	χ_1, χ_2, χ_3	3
Glu	χ_1, χ_2, χ_3	27	Gln	χ_1, χ_2, χ_3	54
Phe	χ_1, χ_2	9	Arg	$\chi_1, \chi_2, \chi_3, \chi_4$	75
Gly	-	1	Ser	χ_1, χ_2	18
His	χ_1, χ_2	12	Thr	χ_1, χ_2	18
Ile	χ_1, χ_2	9	Val	χ_1	3

Lys	$\chi_1, \chi_2, \chi_3, \chi_4$	77	Trp	χ_1, χ_2	12
Leu	χ_1, χ_2	10	Tyr	χ_1, χ_2	18
				TOTAL:	419

The use of rotamers enables computation of side-chain vibrational contributions to the entropy. The calculation of side chain entropy terms is described subsequently in the discussion of the mean field approximation.

The intrinsic vibrational entropy term represents the change from a uniform distribution to the distribution obtained from the partition function described as the equation of Section 5.8.2, and the pairwise vibration entropy term is the change from the initial and main chain-dependent distributions of sub-rotamers of the two side chain rotamers to the distribution of sub-rotamer pairs due to their interaction with each other. Scaling of the side chain vibrational entropy contribution to the pairwise energy by a factor λ is necessary to avoid the overestimation of the entropy change when summing over all pairs of interacting rotamers. This arises for a similar reason to the over-estimation of the solvation energy. In the preferred embodiment of the present invention, λ is given by equation (40):

$$\lambda = \Lambda \frac{N_{\text{first rotamer}}^{\text{contacts}} + N_{\text{second rotamer}}^{\text{contacts}}}{N_{\text{first rotamer}}^{\text{contacts}} N_{\text{second rotamer}}^{\text{contacts}}} \qquad (40)$$

Here, Λ is set to be 0.5.

5.9. ENERGY MINIMIZATION AND ELIMINATION OF INCOMPATIBLE AMINO ACID CONFORMERS

There are two facets of the energy minimization: minimization of the interaction between side chain and template; and minimization of the pairwise interactions between side chains. In either case, there are at least two possible methods of minimization.

5.9.1. METHODS OF ENERGY MINIMIZATION

Side chain conformations taken from the rotamer library of the preferred embodiment of the present invention are idealized. Consequently, in the preferred embodiment of the present invention, their interactions with the protein template and other side chains are optimized and flexibility is introduced to relax strain inherent in the fixed geometries provided by the rotamer library.

Energy minimization is carried out in dihedral angle space using the non-bonding terms of the molecular mechanics force field (van der Waals, electrostatic, and hydrogen bonding).

In one embodiment of the present invention, the energy minimization method may be so-called "Steepest descent"; in another, it may be taken from a class of methods known as "quasi-Newtonian". The theory behind these methods is accessible to one skilled in the art (and can be found in Numerical Recipes in C - The Art of Scientific Computing by WH Press, 2nd edition, section 10.6, SA Teukolsky, WT Vetterling and BP Flannery), but in general terms, the interaction energy and its gradient with respect to displacements in dihedral angle space are utilized. Minima on the energy surface are located iteratively through use of the gradient to search downhill from a given point. The quasi-Newton methods attempt to gather information about the curvature of the energy surface. Methods in this category include "BFGS" and "conjugate gradient", the distinctions between them arising in how each decides to approximate the Hessian (matrix of second derivatives of the energy). In practice, the method of conjugate-gradient has been found to be effective, provided that certain precautionary measures are taken in order to avoid large rotations that would in fact transform one side chain rotamer into another one (this would deteriorate all subsequent partition functions, *i.e.*, that calculated to eliminate the rotamers that have the lowest probabilities according to the interaction energy with the main chain). First, the rotation step size has to be small (fractions of a degree to a few degrees); thus the factors that multiply the gradients are small, and the gradients themselves are capped at some energy maximum. Second, the energy function is modified to contain a rotation penalty function to directly limit the minimization sampling to conformations close to the initial rotamer structure; it has the form in equation (41).

$$\sum_{\text{minimized } \chi} k_\chi (\chi - \chi_o)^2 \qquad (41)$$

The "force constants" k_x are expressed as functions of the standard deviations measured during the construction of the rotamer library by fitting of the frequency distributions observed in the protein database as given by the σ-values in equation (39) for example. Third, if a large rotation (more than a couple of standard deviations) is done despite the penalty, the rotamer is simply placed back in its initial conformation, discarding the result of the minimization.

In a preferred embodiment of the present invention, minimization is carried out by exhaustive sampling of dihedral angles of rotamers close to the ideal conformation. This method is not only simpler, but is superior to conjugate-gradient and similar methods of optimization. (As an example of why this is so, the formation of a hydrogen bond may require an energetically unfavorable rotation of a side chain in order for the correct geometry to be achieved, which would only be discovered using a method that samples rotamer conformations close to the ideal conformation without energy minimization.) The conformations which are obtained through systematic rotations around the dihedral angles χ_1 and χ_2 are referred to as "sub-rotamers". In a preferred embodiment of the present invention, the step size and number of steps are precomputed for each of the twenty naturally occurring amino acids and are determined to cover rotations smaller than two standard deviations away from the minima in dihedral angle space (derived during the creation of the rotamer library). Such a range is usually about 15 degrees, or even smaller than a single standard deviation if it is necessary to optimize the residue set. It is expected that a sequence which enables the packing of rotamers in their ideal conformation (that of the library) should be preferred to another sequence that would necessitate rotations of its side chain rotamers. Although there is an advantage in using many small steps in the thoroughness of coverage, the calculation time has to be considered. In the preferred embodiment, three steps of 5 degrees in each direction around the dihedral angles give good results, i.e., 7 steps in all for each angle. This leads to 49 sub-rotamers, and 49x49 energy calculations to obtain a pairwise energy term.

When considering the interaction of the side chain with the template, the final energy is the weighted average over all possible sub-rotamers. The partition function that defines the weight of state i as a part of a system with N states with energies E_i is defined as follows:

$$w_i = \frac{\exp\left(\frac{-E_i}{RT}\right)}{\sum_{j(including i)}^{N} \exp\left(\frac{-E_j}{RT}\right)} \qquad (42)$$

Accordingly, the contributions of sub-rotamers to the intrinsic energy during minimization are given by equation (42). The E_i comprises molecular mechanics energy terms.

In a preferred embodiment of the present invention, pairs of side chains are minimized using systematic sampling of sub-rotamers. The outcome of using sub-rotamers to optimize pairs of side chains is the weighted average of all possible pairs using a formula such as (43):

$$w_{ij} = \frac{\exp\left(\frac{-E_{ij}}{RT}\right)}{\sum_i \sum_j \exp\left(\frac{-E_{ij}}{RT}\right)} \qquad (43)$$

Minimizing pairs of side chains individually (regardless of other side chains) is assumed to be correct as long as the actual conformations of the side chains depart only slightly from their starting point (to prevent any incompatibility in larger sets of side chains).

5.9.2. PROCESSING RESULTS OF MINIMIZATION

In the method of the present invention, after the intrinsic energy computation, side chain conformers that do not interact favorably with the protein target conformation after minimization are rejected. In one embodiment of the present invention, rotamers for which the intrinsic energy term is above a predetermined threshold are rejected. The use of an

absolute threshold, however, may not be ideal. Intrinsic energies vary in magnitude from site to site because they depend on the actual backbone structure. Poorly resolved structures tend to show many spots with local repulsions. Moreover, strains do exist in proteins where repulsions are common, *e.g.*, in turns, where the tight reversal of the main chain in a turn is often accomplished by placing the ϕ-ψ dihedral angles in less favored regions of the Ramachandran plot. Therefore, in a preferred embodiment of the present invention, the absolute energy threshold is placed high enough (about 50 kcal/mol) to accept enough side chain conformers at every position of the target structure, and a subsequent relative threshold is then applied to keep only the most qualified rotamers. The subsequent relative energy threshold is designed to exclude those rotamers whose partition coefficient weights are very small. The absolute energy threshold is scaled according to the various user-defined weighting coefficients. When all w are 1.0, 50 kcal/mol^{-1} is sufficient, but when all w=0.5, except for solvation, a value between 10 - 20 kcal/mol^{-1}, is appropriate. The subsequent relative threshold is determined by calculating for each amino acid a partition function over all rotamers using the template-side chain interaction terms, equation (42) (i.e., an intrinsic energy term, summed over the remaining rotamers) with the minimum threshold being fixed as a minimal probability of frequency, *e.g.*, 0.001.

The way in which minimization may be applied to other categories of macromolecules will depend upon the complexity of the building blocks and upon the overall structure of the macromolecule. In long extended systems such as nucleic acids, the relevance of the pairwise energy term will be less important. Similarly, in systems with inflexible sidechains or with little conformational flexibility, the need for a thorough minimization protocol will be diminished.

5.10. OPTIMIZATION ROUTINES

For peptides longer than a few residues, an exhaustive sampling of every possible sequence and combination of rotamers is not practical. Hence, in the methods of the present invention, procedures are used which either decrease the size of the sequence space to be covered or rapidly find the probabilities of having particular rotamers at each position of the modeled protein structure. These procedures are often termed "semi-exhaustive".

5.10.1. DEAD-END ELIMINATION

In the method of the present invention, significant reductions of sequence space are obtained by discarding amino acids that cannot belong to the optimal sequence, which is that with the lowest potential energy, and the calculation time is shortened. To eliminate an amino acid, the modeling procedure has simply to exclude all its known side chain conformations. Undesired rotamers are detected by applying the dead-end criterion, which is the underlying principle of the dead-end elimination ("DEE") theorem. The theorem states that, for a given residue i, a particular rotamer, i_r, is not compatible with the global minimum energy conformation ("GMEC") if, for the same residue i, an alternative rotamer i_t exists for which the following inequality holds true (Desmet *et al.*, 1992, *Nature* 356:539-542):

$$E_{i_r}^{\text{template}} + \sum_{j \neq i} \min_s \left(E_{i_r j_s}^{\text{pairwise}} \right) > E_{i_t}^{\text{template}} + \sum_{j \neq i} \max_s \left(E_{i_t j_s}^{\text{pairwise}} \right) \qquad (44)$$

The minimum and maximum functions ($\min_s$ and $\max_s$) cycle over all rotamers j_s of residues j, searching for the rotamer which offers, respectively, the lowest and highest value of the interaction energy with residue i. The rotamers picked by the minimum or maximum function, j_s, do not have to be identical. Thus, the left-hand side of Equation 44 evaluates the best possible interaction of rotamer i_r with the side chains of all other modeled residues, while the right-hand side evaluates the worst possible interaction of an alternative rotamer i_t for the same residue with all the other modeled residues. A side chain rotamer is "dead-ending" if its best interaction with the surroundings is less advantageous than that for another rotamer of the same side chain taken at its worst. Only one rotamer i_t that satisfies the inequality has to be found to qualify i_r as dead-ending.

In another embodiment of the invention, a more powerful version of the elimination criterion is utilized that is less restrictive and therefore more effective (Goldstein, 1994, *Biophys. J.* 66:1335-1340). It states that a side chain rotamer i_r is dead-ending if the energy of the model can be lowered by the choice of an alternative rotamer i_t, while keeping all other side chains fixed. This elimination criterion is described in Equation 45 for the same set of j_s:

$$E_{i_r}^{\text{template}} - E_{i_t}^{\text{template}} + \sum_{j \neq i} \min_s \left(E_{i_r j_s}^{\text{pairwise}} - E_{i_t j_s}^{\text{pairwise}} \right) > 0 \qquad (45)$$

Both dead-end elimination criteria can be extended to pairs of rotamers. The energy contribution of a rotamer pair (i_r,j_s) and its interaction with a third residue/rotamer k_t are given by Equations 46 and 47, respectively:

$$E_{(i_r,j_s)} = E_{i_r}^{\text{template}} + E_{j_s}^{\text{template}} + E_{i_r j_s}^{\text{pairwise}} \qquad (46)$$

$$E_{(i_r,j_s),k_t} = E_{(i_r,k_t)} + E_{(j_s,k_t)} \qquad (i \neq j \neq k) \qquad (47)$$

Then, Equation 44, in one embodiment of the present invention, can be written for pairs of rotamers, as follows:

$$E_{(i_r,j_s)} + \sum_{k \neq i,j} \min_t \left(E_{(i_r,j_s),k_t} \right) > E_{(i_u,j_v)} + \sum_{k \neq i,j} \max_t \left(E_{(i_u,j_v),k_t} \right) \qquad (48)$$

and in a preferred embodiment of the present invention, Equation 43, can be rewritten for pairs of rotamers as follows:

$$E_{(i_r,j_s)} - E_{(i_u,j_v)} + \sum_{k \neq l,j} \min_t \left(E_{(i_r,j_s),k_t} - E_{(i_u,j_v),k_t} \right) > 0 \qquad (49)$$

Dead-ending pairs do not lead to an elimination of a particular amino acid unless one of the participating rotamers is the only possible side chain conformer for the related residue position, in which case the other rotamer of the pair is not compatible with the GMEC and can be discarded. Lasters and co-workers (Lasters *et al.*, 1995, *Protein Eng.* 8:815-822; Lasters and Desmet, 1993, *Protein Eng.* 6:717-722) showed that dead-ending pairs can be safely ignored in the minimum term of Equation 44 and the left-hand term of

the minimum function of Equation 45. Due to the exclusion of dead-ending pairs, the minimum functions might return higher values that strengthen the rotamer elimination criterion.

In the preferred embodiment of the present invention, the dead-end elimination routine follows an iterative process as follows: (a) dead-ending rotamers are eliminated, repeating evaluations of Goldstein's criterion (Equation 45) until no more dead-ending rotamers are found; (b) dead-ending pairs are detected using the first elimination criterion (Equation 44; Desmet *et al.*, 1992, *Nature* 356:539-542) because it is estimated more quickly; and (c) new cycles of rotamer removal as in step (a) are carried out. This continues until no more dead-ending pairs can be found. The more effective but computationally expensive criterion for the detection of dead-ending pairs (Equation 48) is used when many rotamers have been eliminated and the whole cycle is restarted. At the end, if all rotamers of an amino acid at a particular site in the protein are dismissed, sequences containing this particular amino acid at that site are also abandoned.

In one embodiment of the present invention, the DEE routine can be used to determine an optimal set of rotamers for a given sequence. In a preferred embodiment, the routine is not used to limit the output to one optimal set of rotamers, since side chains, particularly those positioned on the solvent-exposed surface, are flexible and adopt different configurations.

5.10.2. MEAN FIELD THEORY

The foregoing will have provided the user with one or more acceptable rotamers for each residue position. This outcome represents a level of statistical uncertainty in the situation being described. It is not adequate to subsequently model the system with merely one rotamer for each residue.

It is desirable to obtain a contribution to the entropic term from all possible conformations of side chains that remain after iterative dead-end elimination has eliminated all sequences that cannot belong to the global energy minimum, as described above in this section. In a preferred embodiment of the present invention, mean field theory (MFT) is utilized to achieve this. MFT is an iterative technique which has found wide application in the physical sciences for describing systems of interacting particles which may adopt many different energy states.

All possible side chain conformations are considered using a mean field approximation that is designed to provide an estimate of the entropy of side chains in both the denatured and the modeled state, *i.e.*, the template. The method attributes to each side chain conformation of all residues in the protein sequence that are not fixed a probability that depends on the average of all possible environments, weighted in turn by their respective probabilities of occurrence (Koehl & Delarue, 1994, *J. Mol. Biol.* 239:249-275; Koehl & Delarue, 1995, *Nat. Struct. Biol.* 2:163-170; Koehl & Delarue, 1996, *Curr. Opin. Struct. Biol.* 6:222-226). The probability of occurrence is related to the energetic favorability. The system is initialized by giving an equal probability to each side chain rotamer so that every side chain conformation "feels" equally the presence of all rotamers at other residue positions. At this point, the field energy of each rotamer is the sum of all possible pairwise interaction energies normalized by the number of interactions, plus a contribution from the interaction with the protein template, as shown in equation (50):

$$MF_{i_r} = E_{i_r}^{\text{template}} + \sum_{\substack{\text{all residues} \\ j}} \sum_{\substack{\text{all rotamers} \\ s}} w_{j_s} E_{i_r j_s}^{\text{pairwise}} \qquad (50)$$

wherein $MF(i_r)$ is the mean field energy experienced by rotamer r of residue i. Initially, the probabilities, w, for the rotamers of any residue are equal. At all times, in order for them to be interpretable as such, these probabilities sum up to 1. The application of MFT is to minimize the term MF by suitable adjustments of the weights.

Having obtained the mean field energy perceived by each rotamer of a residue, proper weights can be estimated from a partition function that integrates all field energies, so that the rotamer that interacts best with its environments becomes more probable than competing rotamers equation (48):

$$w_{i_r} = \frac{\exp\left(\frac{-MF_{i_r}}{RT}\right)}{\sum_{\substack{\text{all rotamers } l \\ \text{of residue } i}}^{N} \exp\left(\frac{-MF_{i_l}}{RT}\right)} \qquad (51)$$

Thus, equation 51 for the weight of a particular rotamer is the partition function that defines the probability of having the rotamer i_r, from a system of N rotamers. R and T are the gas constant and the temperature, respectively.

Several iterations of field energy calculations (now the weighted average) and partitioning are conducted until the set of probabilities is not further modified. It should be clear that this process is "non-linear", i.e., the quantity to be minimized, *MF*, depends on itself through the adjustable quantities, *w*. In such circumstances, convergence may be achieved through one of several methods of non-linear optimization that will be familiar to one skilled in the art. According to Koehle & Delarue (J. Mol. Biol., 239:249-275, (1994) at page 254), convergence is assisted by use of a factor, λ_M, set to 0.50. In a preferred embodiment of the present invention, in order to break out of dead-locked convergence, λ_M can be set according to the formula (52):

$$\lambda_M = 0.5 \pm 0.1 \text{ Rnd}( ) \qquad (52)$$

where Rnd( ) is a random number in the range 0 to 1.0.

In a preferred embodiment of the present invention, it has been found that smooth progress toward the equilibrium state of convergence, is effectively achieved via "simulated annealing". This technique is a computer-based method, which simulates the "heating" of the protein structure to a high temperature followed by "cooling" it. This is done because the high starting temperatures of simulated annealing, *e.g.*, 1000 K, lead to monotonic distributions of probabilities of side chain conformations, thus providing a random and unbiased starting sample of rotamers. The system is then cooled down to sharpen the distributions and optimize the sets of side chain conformations.

The advantage of the mean field method is that the estimated probabilities are similar to frequencies of occurrence, and are coupled to entropy, as shown in the Equation (53).

$$S_i = -R \sum_{\substack{\text{all rotamers} \\ t}} w_{i_t} \ln w_{i_t} \qquad (53)$$

wherein S_i is the side chain conformational entropy of residue *i*, R is the gas constant, and the set of *w* represent the probabilities of occurrence of each rotamer. In a preferred

embodiment, equation (50) is used to compute entropies of a reference state, as in equation (7).

In a preferred embodiment of the present invention, the mean field approximation is used to estimate the weights of all rotamers of the different amino acids in a set of protein fragments obtained from protein structures in the PDB. In another embodiment, amino acids embedded in sample 5-residue extended peptides are employed. From data obtained in this way, the reference entropy of the denatured state can be measured.

The change in entropy of side chains upon protein folding, in both rotamer and sub-rotamer space can therefore be obtained by a comparison of the entropy of the side chain in the native protein, which is then added to the previously determined entropy arising from the fixing of the protein backbone (See Section 5.5) to obtain the total change in entropy upon folding of the protein.

Mean Field Theory is particularly apposite for the study of amino acid sidechains in proteins. In an alternative embodiment, the Mean Field Theory schemes described by Lee (J. Mol. Biol., (1944) 236:918-939) and Lee and Subbiah, (J. Mol. Biol. (1991) 217:373-388), may be employed. Alternative schemes may be employed both for the study of proteins and for applications to other macromolecules. In another embodiment, the iterative scheme used is Monte Carlo sampling.

5.11. RE-EVALUATION OF SOLVATION ENERGIES FOR SEQUENCES WITH LOW SCORING FUNCTIONS

In the preferred embodiment of the present invention, if the scoring function of a sequence, after being stripped of its solvation term, is below a predetermined energy threshold, solvation energies for that sequence are re-evaluated according to two criteria. Otherwise, the sequence is dropped from consideration. Solvation energies obtained in a pairwise manner are removed and re-computed from accessible surface areas derived from the optimized configuration of side chains. In the preferred embodiment, the more detailed solvation parameters of Eisenberg and McLachlan (1986, *Nature* 319:199-203), listed in Table 2, may be used, though other parameter sets would be adequate.

Table 2: Atomic Solvation Parameters

	Atoms	Radii (Å)	Solvation Parameters, σ_i (cal/Å^2)
Hydrophobic Atoms	C	1.9	16
	S	1.8	21
Polar Atoms	N	1.7	-6
	O	1.4	-6
Charged Atoms	N^+	1.7	-50
	O^-	1.4	-24

The accessible surface areas may be measured using the NSC routine (Eisenhaber *et al.* 1995, *J. Comp. Chem.* 16:273-284) or another equivalent method. The result is a more accurate calculation of the potential energy of the mutant protein.

The solvation energy is assessed according to two properties. As with the previous calculation of the energy, the solvation energy of the reference (denatured) state of the protein must be considered. This can be calculated for each amino acid by considering the solvation energy of a reference state. For this purpose, a reference can be obtained from the average solvent-exposure of the amino acid in a 5-residue peptide sequence, as observed in the Protein Data Bank, but without the context of the surrounding protein structure (except for the "capping" residues on the – and C- ends of the sequence. Each residue then behaves as if the sequence were a free chain. It is also necessary to consider the environment of the residue in the protein. For example, exposed hydrophobic residues should be penalized because they may lead to misfolding. Consequently, the solvation energy is calculated by comparing with residues in all the structures in the PDB. By doing this, it is possible to arrive at optimized conformations and sequences for protein-like solvation.

5.12. OUTPUT OF OPTIMAL SEQUENCE RESULTS

After the dead-end elimination procedure of the preferred embodiment of the present invention, many sequences remain; nevertheless, the subsequent steps (see Mean Field

Theory and Refinement, above) ensure that only those conformations and sequences that satisfy predetermined energy thresholds finally surface as candidates for the target structure. The preferred embodiment of the present invention can produce either detailed or limited outputs, depending on the size of the sampled sequence space. In one embodiment, the output is a simple list of sequences and scores (evaluated using the scoring function) that can be sorted according to the calculated potential energy so that the lowest energy sequences may be readily identified. In another embodiment, a more complete output presents a numerical profile of the energy for each sequence produced. The program is also capable of producing a coordinate file (in PDB format) with the structure of the protein having a mutated sequence. If mean field sampling is performed, both the PDB-file and detailed energy outputs correspond to the combination of most probable rotamers. In another embodiment, the detailed energy output includes the effective solution score taking into account all candidate rotamers with the weights they were given, and a second detailed description of the separate pairwise energy terms resulting of the combination of all possible side chain rotamers. If DEE is used for the conformational sampling (without subsequent application of MFT afterwards), then the effective solution score corresponds to the GMEC where one and only one rotamer is retained for each amino acid side chain; the detailed energy file offers nothing else than the separate energy terms which produce the GMEC total energy and the PDB coordinate file represents the GMEC model.

5.13. GENERALIZATION TO NON-PEPTIDES

Whilst the foregoing has focused specifically upon the types of macromolecules known as proteins and their building blocks, amino acids, the methods can readily be applied by one skilled in the art to other categories of macromolecules including, but not limited to, those which can be viewed as comprising a fixed structure attached to which are freely rotating groups. The alternative embodiments which would be required concern the nature of the rotamer library, the choice of reference state, and the property of interest to optimize. Even though amino acid residues have a simplifying feature in that they consist of side chain and a fixed backbone which enables their conformations to be simply expressed as rotamer libraries, other building blocks may also be conveniently modeled by one skilled in the art. Sugar molecules and nucleotide bases have freely rotating groups attached to ring systems, simplifying structural features which would permit the straightforward construction of conformer libraries. Conformers can be obtained from known structures of carbohydrates and nucleic acids respectively or modeled computationally. Published molecular mechanics parameters are based on atom-type only

and therefore in many cases can be utilized in classes of molecules other than those for which they were parameterized.

The idea of a reference state, although usefully expressed as the denatured form when modeling proteins, can be defined differently for other molecules. By analogy with the alanine pentapeptide, a reference saccharide molecule or small sequence of nucleotide bases could be established as a reference structure for carbohydrate or nucleic acid modeling, respectively, in a manner similar to the procedure already described. In other applications it may be useful to utilize an unsolvated molecule as the reference.

Solubility itself may be a property that can be the subject of investigation and optimization with *Perla*.

In applications where the property of interest is an interaction between the target molecule and some binding molecule, the reference state can be considered to be the unbound target molecule or a sum of contributions from the unbound target molecule and unbound binding molecule. This type of application is likely to be widespread, for example: the interaction between DNA and a protein (*e.g.*, a transcription factor); RNA and a protein; the interaction between peptides in solution; the interaction of a polar macromolecule and a lipophilic membrane.

6. EXAMPLE

Example 1

Parts of the SH3 domain of α-spectrin, a small globular protein domain with a β-sheet architecture, were re-designed with *Perla* . Nine residues in the buried core of the domain were replaced by different hydrophobic amino acids. After the evaluation of template - side chain and side chain - side chain interaction energies, the large sequence space was reduced by eliminating dead-ending amino acids. For all remaining sequences, the mean field approximation was used to set the weights of all possible side chain rotamers, providing at the same time an estimation of the change (upon folding) in the entropy of the side chains. The most probable conformation then served for the computation of the change (upon folding) of the protein solvation. A comparison of the best candidates for the replacement of the nine wild-type residues indicated that most of them indeed are constrained by the template and do not tolerate mutations, while others undergo conservative changes. Four residues that form a surface-exposed turn were similarly re-designed, allowing both

nonpolar and polar amino acids. *Perla* produced a unique sequence, also related to the original wild-type sequence.

Choice of template: The three-dimensional architecture (template), residue numbering and wild-type sequence used in the design correspond to the structure presented by Musaccio et al., 1992, (pdb accession code 1shg; Musacchio, A., Noble, M., Pauptit, R., Wierenga, R. and Saraste, M. Crystal structure of a Srchomology 3 (SH3) domain. *Nature*. 359,851-855).

Operating Parameters:

The scoring function: For the various energy terms, recommended weights were set at 1.0. The force field employed was ECEPP.

When calculating molecular mechanics energy terms, an interaction cutoff was employed. (It is well-established that pairwise interactions between atoms separated by greater than a certain distance contribute negligibly.) Here 20 Å is found to be large enough to avoid any important truncation of the electrostatic interaction energy; van der Waals interactions are only taken into account for atom-atom distances smaller than 8 Å.

If the coulomb equation is multiplied by a decaying exponential, the energy reaches zero faster than with the alternative scheme and a smaller cut off might be used. *Perla* selects the screening scheme according to the value of the screening factor. For large values (i.e. more than 1000), a distance-dependent dielectric constant is used. On the contrary, the dielectric constant is fixed and the coulomb equation is multiplied by a decaying exponiential, whose decay factor (1/r,) is set equal to the inverse of the screening factor. The dielectric constants, in our example, are 16 and 4, for the solvent-exposed and buried residues, respectively. Dielectric constants should not be less than half nor more than twice those given in this example, otherwise the importance of the electrostatic term would not be proper (we had a series of good results using constants of 8 and 4 to measure interaction energies between side chains in α-helices).

Modeling sets. In the following examples, *Perla* takes different amino acid side chains from its rotamer library and assembles them op top of nine buried, or four exposed, positions of the SH3 domain. The first modeling set (called CORE) consists of v9, A11, V23, M25, L31, L33, V44, V53 and V58. The second set (called SURFACE) comprises V46, N47, D48 and R49. Since all other side chains are kept in the conformation deposited at the Brookhaven data bank, they constitute the protein template, along with the main chain of the whole protein.

Amino acids considered. For the CORE modeling set, only nonpolar residues (AVILFW) were considered to speed up the sequence sampling, through reducing the total number of sequences. Polar and charged amino acids could be avoided since all residues were to be fully buried. Since there were 9 residues to design, the total number of sequences was 10^7.

For the SURFACE modeling set, both polar and nonpolar amino acids were considered (AVILGDNSTEKRYW); total number of sequences $10^{4.7}$.

For the CORE set, the amino acid considered have 1, 3, 9, 10, 9 and 12 rotamers, respectively, which means that 44 side chains are modeled onto the nine positions, resulting in 396 constructions. A similar calculation indicates that, with a total of 1400 chain conformers, the second design set shows more conformational complexity.

Solvation of the WT protein, used to fix the threshold for good solvation. The solvation energy of the protein target, measured with the sequence and side chain conformations of the wild-type protein, is 4.26 kcal mcl^{-1} (using the parameters of Eisenberg and McLachlan, 1986). This should be compared to the solvation energy state, or some other reference, for a peptide chain of the same sequence composition. Obviously, the amino acids are more exposed to the solvent in the denatured state. Taking as a reference 5-peptides extracted from the protein data bank, to mimic the denatured conformations, the solvation energy of the wild-type sequence is much higher. The difference in solvation energy between the two essential states of the protein folding reaction is thus in favor of the folded three-dimensional structure. This should always be expected, since the solvation energy term is thought to represent the hydrophobic effect that leads to the compaction of protein chains. Besides, taking the average solvation energy calculated in the protein data bank itself, the difference in solvation energy appears as a penalty. That should be the trend for small proteins, for which the buried surface areas are relatively smaller than those of the proteins distributed in the data bank.

Table 3. Original solvation energy (kcal mol^{-1}) (WT sequence with respect to the two possible reference states of the solvation potential

Target	4.26		
		Reference	*Difference*
	5-peptide	32.37	-28.11
	Protein	-1.97	6.23

In addition, for proteins that have a high proportion of turns, the burial of charged amino acids to satisfy some of the main chain hydrogen bond donors is quite common. That is the

case of Glu22, which forms a hydrogen bond to the side chain of the amide group of Ser19. The impact on the solvation energy term is negative.

Reconstruction of side chains, estimation of the template-side chain energy term (van der Waals). Evidently, the hydrogen bonding term plays no role in the design of the CORE set since none of the amino acids considered displays any hydrogen bonding aptitudes, while the electrostatics term has a minor contribution that ensues from the uncharged nature of the hydrophobic amino acids. The major term for this modeling set is the van der Waals term, which Is worth several kcal mol^{-1}. The different behaviors for valines at positions 11 and 44 give an interesting hint about the correlation between the molecular mechanics potential and the rotamer preferences found in the protein data bank. Valine, a β-branched amino acid, populates preferably the extended conformation, matted by values of ψ above 100°, where the optimal side chain rotamer is the trans conformer. That matches the case of position 11 (ϕ=-80 and ψ=120). However, for a value of ψ above 150-160° that coincides to the situation at position 44 (ϕ=-140 and ψ=180°, the most abundant rotamer corresponds to the gauche - conformer. Clearly, the van der Waals energy term measures these different propensities.

Table 4. Interaction with the template (kcal mol^{-1}), modeling set CORE

		Van der Waals	*Electrostatics*	*H-bonding*
Ala 9		-5.78	0.01	-
		-5.77	0.01	-
Ile 9		>50	-0.03	-
		>50	-0.03	-
		5.27	-	-
		-1.24	0.01	-
		-13.25	-	-
		-13.17	-	-
Val 11	gauche -	45.92	-0.10	-
		44.76	-0.11	-
	trans	-11.15	-0.05	-
		-11.21	-0.05	-
	gauche +	42.52	-0.11	-
		39.78	-0.12	-

Val 44	gauche -	-10.21	-0.05	-
		-10.17	-0.05	-
	trans	0.07	-0.05	-
		-0.56	-0.06	-
	gauche +	-2.93	-0.06	-
		-3.46	-0.06	-
Trp 44		161.27	0.03	-
		98.27	0.02	-
		-16.02	-0.04	-
		-18.41	-0.04	-

For each angle for each residue the first and second lines represent the energies prior and after energy optimisation (achieved through the sampling of subrotamers configurations), respectively. Interaction energies of only a few rotamers are shown. Zeros were replaced with - for clarity.

Table 5. Interaction with the template (kcal mol^{-1}), modeling set SURFACE

		Van der Waals	*Electrostatics*	*H-bonding*
8.0[a]				
	Lys 48	-3.71	-0.64	-
		-3.95	-0.77	-
		-4.29	-0.47	-
	Asp 48	-2.68	0.67	-
		-3.08	0.43	-
		-3.70	0.91	-
	Arg 49	-8.92	-3.16	-2.35
		-9.77	-1.06	-
		-8.33	-1.27	-0.58
16.0[b]	*Lys 48*	-3.71	-0.33	-
		-3.95	-0.40	-
		-4.29	-0.24	-
	Asp 48	-2.68	0.34	-

		-3.07	0.22	-
		-3.71	0.47	-
	Arg 49	-8.96	-1.63	-1.91
		-9.76	-0.54	-
		-8.33	-0.65	-0.51

[a,b] Calculations were run using a dielectric constant of 8 or 16. Only the optimised interaction energies are shown, for a few rotamers. Zeros were replaced with - for clarity.

Importance of the reference state. Inspecting the van der Waals interaction energy due to different amino acid side chains, a clear correlation with the size of the chain is found: while the small alanine (one heavy atom) scores about six kcal mol^{-1}, isoleucine (four heavy atoms) contributes approximately 12 and tryptophan (imidazole ring) 18. In the many configurations of the denatured state, contacts between the same side chains and the rest of the protein certainly exist, and a similar scaling should apply. Hence, since we are interested only in the energy difference between denatured and folded states, it is necessary to estimate and remove amino acid type-dependent reference values from the template-related energy term. These are about 3, 5 and 10 kcal mol^{-1} for alanine, isoleucine and tryptophan, respectively.

Reconstruction of side chains, estimation of the template-side chain energy term (H-bonds and electrostatics). To focus on the electrostatic and H-bonding energies, a couple of examples from the second modeling set are given (see table 5). Calculations were performed with different distance-dependent dielectric constants to review the importance of that parameter. Basic amino adds (lysine and arginine) fit better than acidic residues at positions 48 and 49. This is a consequence of the overall negative charge of the protein template in the proximity of the engineered turn. The amplitude of the electrostatic term for residues that are facing the aqueous solvent is the main point of interest, taking into account that water molecules normally screen atomic charges. The electrostatic and hydrogen-bonding interaction energies that correspond to the formation of a salt-bridge between some rotamers of arginine 49 and glutamate 17 of the template, seem to be overestimated when a dielectric constant of eight is used. Besides, replacing aspartate 48 by a lysine would change the electrostatic term by up to 1.7 kcal mol^{-1}, a value that is very likely to be

attenuated by the shielding solvent. Values that are more adequate are obtained with a dielectric constant of 16 or even 32.

Sampling of sub-rotamers to optimize the energy of interaction. In the first example, Table 4, the optimization was achieved by averaging the energy of the rotamer found in the library and other subrotamers, which combine 5° rotations around the bonds that define the χ_1 and χ_2 dihedral angles. In some cases, the energy does increase (e.g. the gauche - conformer of valine at position 44) due to the overall less favorable interaction of the additional subrotamers. For other conformers, slight to significant improvements are obtained (e.g. ~0.53 kcal mol^{-1} for the gauche + of the same residue and -6.51 kcal mol^{-1} for an isoleucine rotamer at position 9). Importantly, large and positive values of the van der Waals term reveal side chain rotamers that are not compatible with the protein template: the Gauche conformers of valine at position 11 and some conformers of isoleucine, position 9 or tryptophan, position 44. The repulsions are not removed with 5° rotations of the side chains. Some caution must be taken, since side chains can deviate by more than five degrees from their ideal conformation, as manifested during the analysis of the protein data bank that led to the generation of the rotamer library.

The optimization should be given more flexibility; a larger range of subrotamers should be sampled. In the following table, some cases are analyzed using two or three 5° steps. The strong repulsions are not removed, even if the algorithm rotates the valine side chain by 15°. The most striking data is that the energy of interaction of the trans and gauche + rotamers, with the template, can be significantly improved if alternative conformations we sampled within the range of dihedral angles observed in the protein data bank. It is essential to decide which scheme of optimization (how many steps, what step size) should be used for the design of a sequence, but it is extremely difficult to assess the quality and validity of the results. Energy changes are important and already in the range of protein stabilities (few kcal mol^{-1}) and shall have a serious impact on the final score of each sequence. The third optimization scheme should be used with care. Indeed, the description of the sequence score by means of template-related energies and side chain pairwise interactions, optimized separately, might lead to a combination of side chains that fit well when taken by pairs, but cannot stand together in the template. The risk for such an occurrence is related to the optimization procedure: the larger the rotations, the higher the risk.

Table 6. Optimisation of the interaction with the template (in local kcal mol^{-1})

			2x5°	
		Van der Waals	*Electrostatics*	*H-bonding*

	Val 11	*gauche-*	43.28	-0.13	-
		trans	-11.25	-0.05	-
		gauche +	36.90	-0.13	-
	Val 44	*gauche -*	-10.14	-0.05	-
		trans	-1.52	-0.06	-
		gauche +	-3.85	-0.06	-
				3x5°	
			Van der Waals	*Electrostatics*	*H-bonding*
	Val 11	*gauche -*	42.07	-0.14	-
		trans	-11.20	-0.05	-
		gauche +	34.42	-0.13	-
	Val 44	*gauche -*	-10.10	-0.05	-
		trans	-2.42	-0.06	-
		gauche +	-4.05	-0.05	-

Main chain entropy loss. When side chain are constructed upon any position, the identity of the engineered amino acid is associated to the ϕ-ψ angles to determine the cost of fixing the backbone, which participates to the entropy change upon protein folding. Although that part of the potential acts similarly to secondary structure propensities, only positive energies are obtained due to the employed mathematical formulation: good secondary structure formers show lower costs (e.g. ~0.7 kcal mol^{-1} for alanine in α-helical conformation and -1.4 kcal mol^{-1} for valine in a β-strand) than breakers *(e.g* ~1.4 kcal mol^{-1} for glycine in α-helical conformation and ~2.9 kcal mol^{-1} in a β-strand). Rarely observed conformations are thought to bring tensions into the protein structure, and are marked by high entropy cost, whatever the amino acid is (see position 47, following table, and by repulsive van der Waals interactions (both terms probably overlap).

Table 7. Interaction with the template (kcal mol^{-1}): entropy costs and reference state

	Energy	*Entropy Cost*	*Reference*	*Sum*
Val 46				
	-10.63			-3.07
	-11.92	2.55	5.01	-4.36
	-10.31			-2.75
Gly 47				

	-0.58	4.76	0.84	5.02
Asn 47				
	14.66			25.16
	13.89			24.40
	49.16			59.67
	22.87	4.71	5.80	33.37
	14.26			24.77
	---			---

Elimination of rotamers not compatible with the template. Obviously, conformers with strong clashes can be removed from the modeling set. We can deduce from the examples shown above that an energy threshold of 0.0 kcal mol^{-1} would probably work fine. To illustrate the limitation of such an absolute threshold, an example will be drawn from the second modeling set, focusing on position 46 and 47. Having a first look at the possible conformations of the wild-type valine at position 46, we see that the template-related energies, including van der Waals, electrostatics, H-bonding and the cost for fixing the main chain dihedral angles, represent favorable interaction energies, even after the reference state contribution is removed. There would be no reason at this stage to eliminate one or the other rotamer, say that which interacts less favorably with the template (third rotamer, actually the gauche + conformer). That same rotamer might be the one that fits better to the target protein structure, If it provides an extra few stabilizing kcal mol^{-1} through the interaction with all other modeled side chains. Consequently, an absolute energy threshold about zero would de excellently since all three rotamers would be maintained.

As shown in the previous table, the template-related energies of an Asn or a Gly residue placed at position 47 follow a distinct trend. For glycine, which has no side chain, the rotamer built by *Perla* corresponds to the second alpha proton. It scores few positive kcal mol^{-1} because the molecular mechanics terms (van der Waals, electrostatics and H-bonding) are largely overwhelmed by the cost of fixing the main chain dihedral angles. At this position, the entropy cost is quite high regardless of the amino acid type (above 4 kcal mol^{-1}), due to the ϕ-ψ dihedral angles that occupy a less-favored region of the Ramachandran plot (ϕ=40 and ψ=-100). All energies related to asparagine rotamers are largely positive, mostly because of repulsions with the main chain atoms. If the energy threshold for elimination were set up at zero, as previously suggested, neither glycine nor

asparagine would be allowed. None of the 20 amino acids would be accepted; *Perla* could not go on with the design. The same absolute threshold simply cannot be used at all positions of the target protein. Neither would it be practical to place position-dependent thresholds, for several reasons. First, the problem just illustrated would not be completely eliminated. Second, there is no simple method to decide what threshold should correspond to which position. Third, different thresholds could introduces position-dependent biases, for example by allowing more or less amino acids types and conformations.

Table 8. Elimination of side chains that are the least compatible with the template

	Energy (kcal mol^{-1})	*Weights*	*Decision*
Asn 47			
	25.16	0.14	
	24.40	0.50	
	59.67	<0.001	dismissed
	33.37	<0.001	dismissed
	24.77	0.26	
	---	---	--
	209.66	<0.001	dismissed
	514.07	<0.001	dismissed
	403.94	<0.001	dismissed
	163.95	<0.001	dismissed
	223.43	<0.001	dismissed

Turning the rotamer-template interaction energies into weights in such a way that information is obtained about an amino acid rotamer with respect to the existing alternative rotamers (through a partition function), elimination is done without taking care about the position-dependent energy ranges. *Perla* has no other choice than keeping the single rotamer of glycine (whose weight is one), while the worst performing conformers of asparagine are dismissed. The absolute threshold is used, still, as a first barrier (50 kcal mol^{-1}) and a lower limit is set for the relative weights (usually 0.001). Both modeling examples were executed with the recommended threshold values, sampling subrotamers either with one, two or three 5° steps around the χ_1 and χ_2 dihedral angles of the library elements. Elimination of side chain conformers than are not (or least) compatible with the template led to a great

reduction of the combinatorial space, therefore, accelerating the stage of side chain - side chain pairwise energy calculations.

Solvation. The presentation of separate intrinsic and pairwise contributions to the solvation energy is not of interest. Solvation energies were computed with values of λ for core and non-core residues, as described above.

Side chain - side chain interaction energies. These are measured and optimized with subrotamers similarly to the way in which the template side chain interaction energies were treated.

Dead-end elimination of sequences. With all possible energy terms stored in memory, *Perla*'s task is to set a score for every sequence by summing over all corresponding terms. Due to the large number of sequences, and the large number of side chain conformations related to each of them, the computational time would be far to long if there was no means to skip uninteresting sequences. The dead-end elimination, a mathematical theorem based on the pairwise formulation of the scoring function, enables the discovery of side chain conformers that cannot belong to the global minimum energy conformation. Such conformers can then be ignored.

Table 9. Dead-end elimination, modeling set CORE

	Number of rotamers		Amino Acids	
Residue Position	Before	After	Before	After
9	9	5	(AVILFW)	AVIL
11	6	3	"	AVI
23	8	4	"	VIL
25	9	6	"	AVIL
31	15	5	"	LI
33	13	3	"	LI
44	8	3	"	AVI
53	9	6	"	AVILF
58	15	6		AVLF
	Number of conformations		Number of sequences	
	$10^{8.9}$	$10^{5.8}$	$10^{7.0}$	$10^{4.5}$

Table 10. Dead-end elimination, modeling set SURFACE

	Number of rotamers		Amino Acids	
Residue Position	Before	After	Before	After
46	61	1	AVILGDNSTEQKRY W	V
47	96	1	"	G
48	153	1	"	S
49	136	1	"	K
	Number of conformations		Number of sequences	
	$10^{9.1}$	1	$10^{4.7}$	1

Mean field approximation, conformational sampling. The mean field approximation sets weights for all existing side chain conformers, depending on the pressure maintained by the surrounding environment (field). Since that exact environment is itself variable, the methodology again costs in an iterative process. Initial weights are established according to the interaction of rotamers with all possible environments given at first equal opportunity. After computing the weights of the rotamers at every position, the interaction between any rotamer with the rest of the modeled side chains is averaged following the weights of their own rotamers. With the new field energies, new weights are obtained. The calculation is repeated until equilibrium, which is indicated by an insignificant variance of the weights. That scheme of cycles if first carried out at high temperature. The system is cooled down after each convergence point, until the temperature used to determine the probability distribution of the reference state is reached. At that stage the energy score of the sequence correspond to a weighted average integrated over all possible side chain conformations. Tables below give some details about the score evolution and entropy change during the simulated annealing run. The entropy of the side chains, derived from the distribution of probabilities, decreases parallel with the temperature decrease. Simultaneously, the fitness score improves due to a larger contribution from low energy conformations (e.g. the GMEC). The number of cycles necessary to reach a stable set of weights increases, possibly

because the energy landscape determined along the multi-dimensional conformation space becomes rougher and rougher.

Table 11. Mean field approximation, modeling set CORE (sequence IVIILLVIV with 2520 conformations)

Temperature (K)	Number of cycles	Absolute entropy (kcal mol^{-1})	Weighted Score (kcal mol^{-1})	GMEC (kcal mol^{-1})
1073	15	9.04	-86.76	
973	23	7.98	-87.04	
573	58	3.88	-88.49	
473	67	2.98	-88.91	-70.77
373	77	2.29	-89.27	
303	86	2.01	-89.42	

Table 12. Mean field approximation, modeling set SURFACE (sequence VGSK with 768 conformations)

Temperature (K)	Number of cycles	Absolute entropy (kcal mol^{-1})	Weighted Score (kcal mol^{-1})	GMEC (kcal mol^{-1})
1073	16	12.60	5.11	
973	23	11.26	4.94	
873	30	9.93	4.75	
473	63	4.56	3.66	2.06
373	72	3.26	3.28	
303	81	2.38	2.99	

Thresholds for elimination of sequences. Reconstructing the nine wild-type residues corresponding to the CORE modeling set (VAVMLLWV), the sequence-to-structure score excluding the solvation term is -48.47 kcal mol^{-1}. Hence a threshold of -40.0 kcal mol^{-1} would offer a good first selection level. The solvation term then contributes about the same amount of energy as the original X-ray structure, pushing the fitness score below -70 kcal mol^{-1}. For the second modeling set, the first energy score is positive, due to the strain related to the main chain configuration of the turn. The solvation term was slightly improved thanks

to the side chain placement carried out by *Perla*. To design the turn sequences, the two energy thresholds can be set to 30.0 and 0.0 kcal mol^{-1}, respectively.

Table 13. Sequence-to-structure fitness score (kcal mol^{-1})

	Input Target	Wild-type Sequence	
		Modeling set CORE	Modeling set SURFACE
Energy		-48.47	24.30
Solvation term	-28.11	-28.23	-30.91
Total		-76.70	-6.61

Score of the sequences. Many possible sequences had to be sampled for the CORE design example. When such a case is expected, it is convenient to simplify the output data written by *Perla*, and ask the program to produce a single output file with a list of sequences and fitness scores, rather than many detailed energy files. Also, there is no need to have a PDB file for every sequence, since these could not be examined. The program can be executed afterwards with a reduced set of amino acids or more stringent acceptance threshold levels, in order to obtain the fully detailed output. In figure 13, the fitness score is plotted versus the ordered list of sequences designed for the CORE

Designed sequences. Sequences engineered by *Perla* resemble the wild-type (WT) sequence, for both design examples. Out of nine residues, four amino acids were preserved in the CORE set (best sequence IVIILLVIV). Only one out of four is maintained for the SURFACE example, but the Asp48 to serine and Arg48 to lysine mutations are conservative (unique designed sequence VGSK).

Output Sequences. Sample sequences along with their solution scores as output from the program are as follows:

Table 14

CORE			SURFACE		
VAVMLLVVV	-76.6	(Wild Type)	VNDR	-6.6	(Wild Type)
LVIVLLVIV	-81.8		VGSK	-28.0	
VVIILLVIV	-81.8				
IVIILLVVV	-81.9				

LIIVLLVIV	-82.0
IVVILLVIV	-82.8
IIIVLLVIV	-82.8
IVLILLVIV	-82.9
LVIILLVIV	-83.2
IVIILLVIV	-84.4

We note that the solution scores of the best solutions are all somewhat lower than the score of the "wild type" sequence. Figure 13 shows the distribution of solution scores for approximately 1600 considered solutions.

EXAMPLE 2. Computer-assisted re-design of the SH3 domain and experimental characterization of the mutant proteins

Perla was used to re-design the nonpolar core of the SH3 domain of *a*-spectrin, (Musacchio, A., Noble, M., Pauptit, R., Wierenga, R. and Saraste, M. (1992) Crystal structure of a Src-homology 3 (SH3) domain. *Nature*, **359**, 851-855), and four solvent-exposed turns. Sequences engineered by the design algorithm could be interpreted in terms of packing optimization, high secondary structure propensities, and favorable hydrogen bonding and electrostatics interactions. Protein mutants were produced and characterized using circular dichroism, and urea-induced equilibrium unfolding was monitored by fluorescence to determine the relative protein stabilities. Most mutants do fold to the desired target conformation, and some have higher stabilities than the wild-type protein.

These protein design applications were based on an early version of the computer program that used only the dead-end elimination principle as a basis for sampling (the technique thus resembled that of Dahiyat, B.I. and Mayo, S.L. (1996) Protein design automation and *Protein Sci*, 5, 895-903 and Dahiyat, B.I., Gordon, D.B. and Mayo, S.L. (1997)). Automated design of the surface positions of protein helices. *Protein Sci*, 6, 1333-1337. Hence, a unique conformation, that of minimal energy, was associated with each sequence, and there was no side chain entropy term. The optimization of interaction energies was based on conjugate-gradient minimisation instead of subrotamers sampling. Solvation was not yet accounted for

in the pairwise description of the energy function. Nonetheless, the calculation processes were similar to what was described in Example 1: sequence designs were conducted fixing all side chains in the conformation of the X-ray template structure, except for the residues to be replaced, and similar choices of amino acids were considered.

5

Re-design of the buried protein core

To replace the nine residues buried into the protein core, we let *Perla* choose Ala, Val, Ile, Leu, Phe or Trp. An example corresponding to this computational problem was presented in EXAMPLE 1: the optimal candidate sequence was IVIILLVIV, which contains five mutations with respect to the wild-type sequence (VAVMLLVVV). Using the early version of *Perla*, we picked up a somewhat different sequence, VVLILLVIL, which also contains five mutations. Many other sequences were proposed (the dead-end elimination of non-optimal amino acids being not totally efficient). As for the example in EXAMPLE 1, amino acid elimination was not equal for the nine residue positions (Table 15). Only tryptophan was eliminated at all positions, probably because its larger side chain could not fit into the spatially restricted internal region of this SH3 protein domain.

20

Table 15. Amino acids participating to the most optimal sequences (Core cluster)

Position	*Wild-type residue*	*Initial choice*	*Optimal sequences*
9	Val	[A,V,I,L,F,W]	A,V,I,L
11	Ala	"	V,I,L
23	Val	"	V,I,L
25	Met	"	A,I,L
31	Leu	"	A,V,I,L
33	Leu	"	A,I,L
44	Val	"	A,V,I,L
53	Val	"	A,V,I,L,F
58	Val	"	V,I,L

30

The energy thresholds that help reduce the number of sequences given by the design algorithm limited the output to approximately 3000 sequences (out of the 10^7 sequences initially possible). We decided to construct two protein mutants, relying on the sequence with optimal sequence-to-structure relationship and on another with less optimal score. This second sequence (VVLLLLAFL) was selected because it provided one more mutation and contained an aromatic residue (Phe 53) whose larger side chain produces a higher impact on the geometrical organization of the neighboring side chains. Only ~1.5% of the candidate sequences contained

a Phe at position 53, all of them being correlated with the presence of Ala in the facing position 44. When comparing the SH3 domain X-ray structure with the structures modeled by Perla, for these two new core sequences, we observed that the packing density is increased from wild-type to first mutant to second mutant.

Re-design of the turns

Four turns were designed independently, replacing in each the four wild-type residues by amino acids taken in the following list: A, G, V, I, L, D, E, K, R, H, S, T, N, Q, W and Y (total number of sequences: 65536). An example corresponding to the modeling of residues 46 to 49 (distal loop) was presented in EXAMPLE 1. Using a version of *Perla*, we obtained a single optimal sequence for three of the turns (diverging turn, n-Src and distal loop of the SH3 domain), while 18 candidate sequences were possible for the fourth one (RT-loop). Table 16 presents the designed sequences.

Table 16. Designed mutants for the four turns.

Cluster (residues)	*Wild-type sequence*	*Designed sequence*
RT-loop (19-22)	S,P,R,E	R,S,D,E
diverging turn (26-29)	K,K,G,D	R,H,G,D
n-Src loop (38-41)	N,K,D,W	D,K,D,R
distal loop (46-49)	V,N,D,R	I,G,T,K

Experimental characterization of the protein mutants

Re-designed proteins were produced by means of recombinant DNA technology and molecular biology techniques. Protein expression was not especially high, yet protein yields after purification were sufficient to analyze all protein mutants except that corresponding to the diverging turn (residues 26-29) and n-Src loop (residues 38-41). To check that the proteins were correctly folded, far UV CD spectra were recorded. Equilibrium unfolding transitions were examined to evaluate the protein stabilities.

Structural characterization via far-UV CD spectroscopy

Core mutants. The shape of the far UV CD spectrum of the α-spectrin SH3 domain is unusual: it does not show the typical signal of a β-sheet containing protein (a minimum around 217nm) but displays two maxima at approximately 220nm and 235nm, separated by a minimum at 227nm (Figure 14, A). These specific peaks, attributed to the two consecutive tryptophan amino acids (W41-W42), give an indication that the correct tertiary structure is formed; they disappear when the protein unfolds (see the high temperature spectrum in Figure 14, A). Hence, we observe that the first core mutant is correctly folded (Figure 14, B). The second mutant, however, is nearly unfolded at 298 K but folds properly at lower temperature, attesting that it is less stable than the WT SH3 domain (Figure 14, C).

Turn mutants. Similarly, the three designed mutants we were able to produce for the turn clusters (RT-loop, diverging turn and distal loop) are correctly folded (Figure 15).

Urea-induced equilibrium unfolding

To compare protein stabilities, we have monitored the equilibrium unfolding, induced by urea, recording the change in emitted fluorescence. The unfolding curves were fitted with Equation (54) (based on the linear extrapolation method; Pace, C.N. (1986) Determination and analysis of urea and guanidine hydrochloride denaturation curves *Methods Enzymol*, **131**, 266-280; Santoro, M.M. and Bolen, D.W. (1988) Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants. *Biochemistry*, **27**, 8063-8068), which corresponds to a two-state unfolding mechanism but has been improved to include a special dependence of the fluorescence of SH3 proteins on urea concentration. In equation (54), the parameters of interest are the slope in the transition region ($m_{u\to f}$) and the folding free energy in the absence of denaturant ($\Delta G_{u\to f}^{H_2O}$). Figures 16 and 17 show the unfolding curves and the parameters obtained from data fitting to Equation (54) are summarized in Table. Data fitting was improved defining explicitly the slope in the transition region to –0.93 (the average of the values corresponding to the wild-type protein), as observed by the smaller difference between folding free energies obtained from separate experiments.

$$F_{obs} = \frac{\left(F_f + m_f[D]\right) + \left(F_u + m_u[D]\right)\exp\left[-\left(-\Delta G_{u\to f}^{H2O} + m_{u\to f}[D] + a[D]^2\right)/RT\right]}{1 + \exp\left[-\left(-\Delta G_{u\to f}^{H2O} + m[D] + a[D]^2\right)/RT\right]}$$

(54)

Equation (54) is an expression of the observed fluorescence (F_{obs}) as a function of the denaturant concentration ([D]), for proteins that have a two-state unfolding transition. The linear dependence on denaturant concentration, of the fluorescence of the folded (F_f) and unfolded (F_u) states is taken into account, with slopes m_f and m_u, respectively. The protein stability or folding free in energy in absence of denaturant is $\Delta G_{u\rightarrow f}^{H_2O}$ and $m_{u\rightarrow f}$ is the slope in the transition region. The quadratic term, particular to the SH3 domain of a-spectrin Prieto, J., Wilmans, M., Jimenez, M.A., Rico, M. and Serrano, L. (1997) Non-native local interactions in protein folding and stability: introducing a helical tendency in the all beta-sheet alpha-spectrin SH3 domain. *J Mol Biol*, **268**, 760-778; Viguera, A.R., Serrano, L. and Wilmanns, M. (1996) Different folding transition states may result in the same native structure. *Nat Struct Biol*, 3, 874-880, is added to improve data fitting ($\alpha = 0.008925$). R and T are the gas constant and temperature, respectively.

Core mutants. The unfolding curves in Figure 16 clearly indicate that the first core mutant is more stable than the wild-type SH3 domain, while the other mutant is much less stable. Data fitting to Equation (54) produced parameters in agreement with published data for the wild-type protein (Table 17). For the first core mutant, stability is increased by 0.7 kcal mol^{-1}. For the second mutant, the fitting gave an indication of the amount of folding free energy lost (at least 2.5 kcal mol^{-1}), but the absence of the initial part of the curve renders the modeled parameters quite imprecise. Slopes in the unfolding transition regions are similar, which suggests that the hydrophobic surface areas exposed to solvent (in the unfolded and folded states) have not changed significantly.

Turn mutants. The n-Src loop protein mutant was not successfully purified and the diverging turn mutant was not purified in sufficient quantities, hence, only two of our designed proteins for the turn clusters were characterized (Figure 17). One of them, the RT-loop mutant, is just slightly more stable than the wild-type SH3 domain. On the other hand, modifying the distal loop residues resulted in a considerably larger stabilization. Slopes in the unfolding transition regions were again similar to that of the wild-type protein domain.

Table 17. Folding free energies of the SH3 domain from α-spectrin and mutants, and prediction by Perla of the free energy changes due to the sequence mutations.

Protein	$m_{u\to f}$ [a]	$m_{u\to f}$	$\Delta G^{H_2O}_{u\to f}$ [a]	$\Delta G^{H_2O}_{u\to f}$ [b]	$\Delta G^{H_2O}_{u\to f}$ [c]	$\Delta G^{H_2O}_{u\to f}$ [d]	$\Delta\Delta E_{Perla}$
Distal loop	–0.94	–0.81	–3.86	–3.54	–3.82	–3.85	–8.95
Core 1st	–0.89	–0.96	–3.49	–3.66	–3.55	–3.57	–7.03
RT-loop	–0.85	–0.77	–2.81	–2.58	–3.06	–3.10	–0.83
Wild-type	–0.88	–0.98	–2.72	–3.02	–2.84	–2.85	–
Core 2nd	ND	–0.96	ND	–0.24	– 0.0 [f]	–0.02	–4.45
diverging turn	insufficient protein yield after expression and purification						1.8
n-Src loop	insufficient protein yield after expression and purification						2.48

[a,b] Parameters obtained from data fitting to Equation 54 and [c,d] fixing the slope in the transition region to –0.93; [a,c and b,d] first and second experiments. [e] Difference in scoring energy (mutant – WT). The current version of *Perla* was used to model the wild-type and designed sequences, building the side chains of all residues and with 0.5 weights for all molecular mechanics energy terms and entropy changes, distance-dependent dielectric constants of 8 and 4 for solvent-exposed and buried regions of the protein, 2 steps of 5° rotations to generate subrotamers and a sampling temperature of 303K. (ND) Not determined (correct data fitting not possible).

EXPERIMENTAL PROTOCOLS, METHODS

Gene construction and cloning. Nucleic acid sequences were obtained by reverse-translating (using the preferred codon usage of *E. coli*) the protein sequences, either wild-type SH3 domain of *a*-spectrin or the designed mutants. Full-length genes were engineered, thus including the region coding for the N-terminal first 5 residues (MDETG, including the initial methionine), which are not observed in the X–ray structures of previously characterized mutants and wild-type. All DNA sequences were built using a polymerase chain reaction (PCR) method with oligonucleotides synthesized by the EMBL DNA service. Two central oligonucleotides were annealed together and polymerization was accomplished to obtain a double-stranded DNA fragment, which was further elongated and amplified using two external primers. Two stop codons (TAG) were introduced at the 3' end of the DNA sequence. NcoI (CCATGG) and HindIII (AAGCTT) restriction sites were designed at the 5' and 3' termini of the produced

DNA, respectively, to allow cloning into pBAT-4 (Paränen, J., Rikkonen, M., Hyvönen, M. and Kääriäinen, L. (1996) T7 vectors with modified T7lac promoter for expression of proteins in Escherichia coli. *Analytical Biochemistry*, 236, 371-373) after digestion with the corresponding restriction enzymes. Electro-competent or chemically-competent XL-1 Blue *E. coli* cells were transformed with the engineered plasmid vectors, and the cells were grown on L-broth plate containing ampicilin. Positive clones were selected after screening by PCR (with primers complementary to the vector at the 5' and 3' sides of the cloning site) for bacterial templates that generate a DNA fragment of the expected size (about 200 base pairs). Gene sequences were confirmed by chemical sequencing of the vector cloning site, performed following the dideoxy-mediated chain termination method of Sanger, (Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74, 5463-5467) after purification of the vectors from these positive clones.

Protein expression, extraction and purification. Electro-competent or chemically-competent *E coli* BL21 (DE3) cells were transformed with the purified expression vectors and grown at 37°C (6l scale) from a single colony in L-broth medium containing 50mg l^{-1} ampicilin until the culture reached an optical density of ~0.6 at 600nm. Isopropyl-b-D-thiogalactopyranoside (IPTG) was then added to a final concentration of 40mg l^{-1} to induce the expression of the engineered protein. Cells were harvested three hours later by centrifugation (about 3300 RPM for half an hour) and resuspended in 10mM sodium citrate pH 3.5 and 100 mM NaCl, lysed by sonication and ultracentrifuged at 37500 RPM for 2h. The resulting pellets (containing insoluble proteins) were resuspended in 6M urea, 10mM sodium citrate pH 3.5 and 100 mM NaCl, while polyethyleneimine (PEI) was added to the soluble fractions (containing soluble proteins) to precipitate DNA fragments. Both fractions were submitted to a second run of ultracentrifugation. Electrophoresis on sodium dodecyl sulfate-polyacrylamide gels (SDS-PAGE) was used to determine which fractions contained the overexpressed proteins. While proteins from the insoluble fractions were directly purified, proteins present in the soluble fractions were first precipitated with ammonium sulfate in two steps. Ammonium sulfate was added to the solution to reach a concentration of 30% and the sample was ultracentrifuged (at 25000 RPM for half an hour). Some additional amount of ammonium sulfate was then added to the soluble part to increase the concentration to 70% and a new centrifugation run performed. Pellets were as before resuspended in 6M urea, 10mM sodium citrate pH 3.5 and 100 mM NaCl, and the suspensions submitted to ultracentrifugation. Proteins purification was carried out by exclusion chromatography in denaturing conditions (6M urea, 10mM sodium citrate pH 3.5 and 100 mM NaCl) on a HiLoad 26/60 Superdex 75 column. Fractions with pure protein

were detected with SDS-PAGE, pooled together and diluted 10 times (in 10mM sodium citrate pH 3.5 and 100 mM NaCl) to allow refolding, and were afterwards concentrated by ultrafiltration. The purification was repeated once and purity checked in polyacrylamide gels (purity estimated to be >95%). After refolding and concentration, protein solutions were dialyzed against water at pH3.5. Purity and protein identity were checked by mass spectroscopy by the EMBL peptide & protein service. About 10-20mg pure proteins were obtained.

Determination of protein concentration. Protein concentrations were determined using the method of Gill, S.C. and von Hippel, P.H. (1989) Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem*, 182, 319-326. Absorbance (A_{280}) of tyrosine and tryptophan residues was measured at 280nm, for folded and unfolded polypeptide samples. The folded samples were buffered aqueous solutions of the protein and the unfolded samples similar solutions containing additionally 6M guanidinium hydrochloride (GdnHCl). The following equations (55 and 56) were used to determine the extinction coefficient at 280nm (ε_{280}) and the polypeptide concentration, respectively (with l the cell path length, usually 1cm).

$$\varepsilon_{280}^{folded}\left(M^{-1}cm^{-1}\right) = \frac{A_{280}^{folded}}{A_{280}^{unfolded}}\varepsilon_{280}^{folded} = \frac{A_{280}^{folded}}{A_{280}^{unfolded}}\left[N_{Tyr}1280 + N_{Trp}5690\right] \quad (55)$$

$$[peptide](M) = \frac{A_{280}}{\varepsilon_{280} l} \quad (56)$$

Far-UV circular dichroism. CD spectra were recorded on a Jasco-710 instrument calibrated using D-10-camphorsulfonic acid. Measurements were made every 0.1nm, with a response time of 1s and a bandwidth of 1nm, at a scan speed of 50nm min^{-1}. Protein concentrations were about 100mM; samples were not buffered, the pH being adjusted with HCl or NaOH to 3.5, and temperature was set as indicated in the figures. CD spectra were recorded using a cuvette with a 0.2mm path. Spectra shown in the text are the average of 20 scans, which were corrected for the baseline signal.

Urea-induced equilibrium denaturation. Urea titrations were realized using a Jasco-710 instrument equipped with an automated titration system. A 2ml sample (about 15mM protein in 50mM sodium citrate pH3.5 and 50mM sodium chloride) was placed in a cuvette with a path length of 1cm. Two software-controlled syringes were used to replace in a stepwise manner a fix volume of the sample with a urea-containing protein solution (15mM protein in 50mM sodium citrate pH3.5, 50mM sodium chloride, an at least 8M urea). Thus, the urea concentration was increased while the buffer and protein concentrations were kept constant. 20 urea injections of 50ml were performed first and followed by 15 injections of 100ml, in order to perform measurements for urea concentrations between 0M and 6M (the protocol being limited by the 2.5ml volume of the syringes). After each urea injection, the sample was allowed to equilibrate for 2 minutes, and the total fluorescence above 340nm was recorded (excitation wavelength was 280nm). Constant stirring of the solution was used to facilitate sample equilibration. Temperature was maintained at 298K. For the urea-containing protein solution and the final protein sample, urea concentrations were determined from refractive index measurements as indicated by Equation 57 (where ΔN is the difference between the refractive index of the denaturant-containing sample and the refractive index of the denaturant-free buffer; Pace, C.N. (1986) Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol*, 131, 266-280. Denaturant concentrations for each measurement were calculated according to the experiment protocol and the urea concentration of the urea-containing protein solution used for urea additions. The difference between the measured final denaturant concentration and the concentration expected due to the unfolding protocol was less than 0.2M. The stability of pH was also checked and confirmed at the end of the experiment.

$$[urea](M) = 117.66(\Delta N) + 29.753(\Delta N)^2 + 185.56(\Delta N)^3 \quad (57)$$

7. REFERENCES CITED

All references cited herein are incorporated herein by reference in their entirety and for all purposes to the same extent as if each individual publication or patent or patent application was specifically and individually indicated to be incorporated by reference in its entirety for all purposes.

Many modifications and variations of this invention can be made without departing from its spirit and scope, as will be apparent to those skilled in the art. The specific embodiments described herein are offered by way of example only, and the invention is to be limited only by the terms of the appended claims, along with the full scope of equivalents to which such claims are entitled.

What is Claimed is:

1. A method for choosing a set of substitute building blocks for a set of positions in a target macromolecule according to whether a calculated solution score is lower than a threshold value, the method comprising:

(a) specifying at least one substitute building block for each position in said set of positions to produce a specified set of substitute building blocks;

(b) for each said substitute building block,

i) determining at least one candidate conformer;

ii) substituting coordinates of each said candidate conformer or portion thereof for coordinates of the building block or portion thereof at said position in an atomic structure of said target macromolecule; and

(c) minimizing the value of a calculated energy term by adjusting the geometry of each said candidate conformer or portion thereof in order to obtain a solution structure;

(d) calculating a solution score for said solution structure, wherein said solution score comprises an entropic term; and

(e) choosing said specified set of substitute building blocks if said calculated solution score is lower than a threshold value.

2. The method of Claim 1 wherein said macromolecule is a peptide or protein; said building blocks are amino acid residues; and each candidate conformer is a side chain rotamer selected from a plurality of side chain rotamers.

3. The method of Claim 2 wherein said calculated solution score comprises a difference between a first value corresponding to said solution structure and a second value corresponding to a reference structure.

4. The method of Claim 3, wherein said first value corresponding to said solution structure accounts for interactions between said side chain rotamer and said atomic structure, and a sum of interactions between all pairs of all possible side chain rotamers.

5. The method of Claim 3, wherein said reference structure is a denatured state of said solution structure.

6. The method of Claim 4, further comprising a step of rejecting a side chain rotamer when the value of said interactions between said side chain rotamer and said atomic structure is greater than a threshold value.

7. The method of Claim 2, wherein the dihedral angles of said side chain rotamers are optimized in step (c).

8. The method of Claim 2, wherein the positions of all main chain atoms of said atomic structure, and the positions of all atoms in amino acid side chains that are not included in said set of substitute building blocks are held fixed in said atomic structure.

9. The method of Claim 7, wherein the positions of all atoms in amino acid side chains on residues that are not at said set of positions are allowed to vary whilst the dihedral angles of said rotamer are optimized.

10. The method of Claim 7, wherein the positions of all main chain atoms of said atomic structure are allowed to vary whilst the dihedral angles of said rotamer are optimized.

11. The method of Claim 2, wherein said plurality of side chain rotamers is a library of predetermined rotamer conformations.

12. The method of Claim 2, wherein said plurality of side chain rotamers is derived from a continuous distribution of conformations.

13. The method of Claim 1, wherein

said atomic structure includes a representation of the building blocks at each position in said set of positions; and

said atomic structure was determined by a method selected from the group consisting of x-ray crystallography, nuclear magnetic resonance spectroscopy, electron microscopy, homology modeling, and *ab initio* molecular modeling.

14. The method of Claim 1, wherein said atomic structure is an X-ray crystal structure of a portion of said macromolecule that comprises said building blocks at each position.

15. The method of Claim 14, wherein said X-ray crystal structure was determined at a resolution of less than 4.0 Angstroms.

16. The method of Claim 2, wherein said calculated solution score is calculated using an empirical scoring function.

17. The method of Claim 16, wherein said empirical scoring function is a sum of energy terms, comprising a template energy of said atomic structure held in a fixed geometry, an intrinsic energy of interaction between a candidate side chain rotamer and said atomic structure held in a fixed geometry and a pairwise energy of interaction between possible pairs of side chain rotamers in said substitute set of building blocks.

18. The method of Claim 17 wherein said intrinsic energy of interaction is computed as either $E_{\text{fixed structure, best side chain}(i)}$ or $\sum_{\substack{\text{rotamers} \\ r}} w_{i,r} E_{\text{fixed structure, side chain}(i,r)}$

wherein $E_{\text{fixed structure, side chain}(i,r)}$ is an energy of interaction between the atomic structure held in a fixed geometry and rotamer r of side chain i, and $w_{i,r}$ is a weighting factor.

19. The method of Claim 17, wherein said pairwise energy of interaction is computed as either $E_{\text{best side chain }(i),\text{ best side chain}(j)}$ or $\sum_{\substack{\text{rotamers } r \\ \text{of residue } i}} \sum_{\substack{\text{rotamers } s \\ \text{of residue } j}} w_{i,r} w_{j,s} E_{\text{side chain}(i,r),\text{ side chain}(j,s)}$ wherein $E_{\text{best side chain }(i),\text{ best side chain}(j)}$ is the energy of interaction between side chain rotamer i and side chain rotamer j, and $w_{i,r}$ and $w_{j,s}$ are weights.

20. The method of Claim 17, wherein the template energy of said atomic structure held in a fixed geometry comprises at least one energy term selected from the group consisting of a molecular mechanics potential, a solvation energy, an empirical penalty function, and an entropic contribution.

21. The method of Claim 20, wherein the template energy of said atomic structure held in a fixed geometry comprises a sum of terms whose coefficients are individually adjustable weighting factors.

22. The method of Claim 20, wherein said molecular mechanics potential comprises at least one energy term selected from the group consisting of bond length vibrations, bond angle bends, the hydrogen bond energy between pairs of hydrogen bond donor and acceptor atoms, an electrostatic interaction energy between pairs of charged atoms, and a van der Waals interaction energy between pairs of non-bonded atoms in said atomic structure.

23. The method of Claim 22, wherein the van der Waals energy term is expressed as

$$E_{vdw} = \sum_{\text{nonbonded } ij} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right)$$

wherein the sum runs over all possible non-bonded atom pairs i and j from said atomic structure held in a fixed geometry.

24. The method of Claim 22, wherein the hydrogen bonding energy between pairs of hydrogen bond donor and acceptor atoms is expressed as

$$E_{HB} = \sum_{\text{H-bonded } H,A} \left(\frac{A_{HA}}{r_{ij}^{12}} - \frac{B_{HA}}{r_{ij}^{10}} \right)$$

wherein the sum runs over all hydrogen bonds in said atomic structure held in a fixed geometry and atoms i and j are donor and acceptor atoms participating in each of said hydrogen bonds.

25. The method of Claim 22, wherein said electrostatic interaction energy between pairs of charged atoms is expressed as

$$E_{elec} = \sum_{\text{charges } ij} \frac{q_i q_j e^2}{4\pi \varepsilon_0 \varepsilon_r r_{ij}}$$

wherein the sum runs over all pairs of charged atoms i, and j in said atomic structure held in a fixed geometry whose respective charges are q_i and q_j.

26. The method of Claim 20, wherein said entropic contribution comprises at least one term selected from the group consisting of a main chain entropy term, a side chain rotation entropy term and a side chain vibration entropy term.

27. The method of Claim 26 wherein said main chain entropy term is calculated as

$$-w_{\text{main chain}}^{\text{entropy}} RT_{phys} \sum_{\text{all residues } i} \ln \frac{\sum_{\substack{\text{subspaces } \phi\psi_{20^\circ \times 20^\circ} \\ \text{close to } \phi_i,\psi_i}} w_{\phi\psi_{20^\circ \times 20^\circ}} N_{\phi\psi_{20^\circ \times 20^\circ}}^{\text{amino acid } i}}{N_{\text{all } \phi\psi}^{\text{amino acid } i}}$$

wherein $w_{\text{main chain}}^{\text{entropy}}$ is a coefficient, T is temperature, and N is the number of amino acids of a particular type in a database of known protein structures that are found within a specific range of ϕ, ψ angles.

28. The method of Claim 26 wherein said side chain rotation entropy term is calculated as

$$-w_{\text{side chain}}^{\text{entropy}} T_{phys} \sum_{\text{all residues } i} \left(\left(-R \sum_{\substack{\text{all rotamers } r \\ \text{of residue } i}} w_r \ln w_r \right)_{\substack{\text{target} \\ \text{structure}}} - \left(-R \sum_{\substack{\text{all rotamers } r \\ \text{of residue } i}} w_r \ln w_r \right)_{\substack{\text{reference} \\ \text{structure}}} \right)$$

wherein $w_{\text{side chain}}^{\text{entropy}}$ is a coefficient, T_{phys} is a temperature, and w_r is obtained from a partition function.

29. The method of Claim 26, wherein said side chain vibration entropy term is calculated as:

$$-w_{\text{side chain}}^{\text{vibration}} T_{phys} \sum_{\text{all residues } i} \sum_{\substack{\text{all rotamers } r \\ \text{of residue } i}} w_r \left(\left(-R \sum_{\substack{\text{all sub-rotamers } s \\ \text{of rotamer } r}} w_s \ln w_s \right)_{\substack{\text{target} \\ \text{structure}}} - \left(-R \sum_{\substack{\text{all sub-rotamers } s \\ \text{of rotamer } r}} w_s \ln w_s \right)_{\substack{\text{reference} \\ \text{structure}}} \right)$$

wherein $w_{\text{side chain}}^{\text{vibration}}$ is a coefficient, w_r is a weight, and w_s is obtained from a partition function.

30. The method of Claim 20, wherein said solvation energy is calculated as

$$w^{\text{solvation}} \left(\left(\sum_{\text{atoms } i} \sigma_i ASA_i \right)_{\substack{\text{target} \\ \text{structure}}} - \left(\sum_{\text{atoms } i} \sigma_i ASA_i \right)_{\substack{\text{reference} \\ \text{structure}}} \right)$$

wherein σ_i is a parameter specific to atom i, $w^{\text{solvation}}$ is a coefficient, and ASA_i is an accessible surface area of atom i and atom i is in said atomic structure.

31. The method of Claim 30, wherein said parameter specific to atom i reflects the properties of a solvent selected from the group consisting of water and an organic solvent.

32. The method of Claim 31 wherein the organic solvent is selected from the group consisting of methanol, methylamine, and dimethyl sulphoxide.

33. The method of Claim 20, wherein said empirical penalty function is calculated as

$$- w^{stat} RT_{stat} \sum_{\text{all residues } i} \ln P^{stat}_{\text{amino acid } i}$$

wherein $P^{stat}_{\text{amino acid } i}$ is a term representing a probability of occurrence of amino acid i in nature, T_{stat} is temperature, and w^{stat} is a coefficient.

34. The method of Claim 5, wherein said reference structure comprises said side chain rotamer substituted for a side chain in an alanine based penta-peptide.

35. The method of Claim 5 wherein said reference structure comprises said side chain rotamer embedded in a fragment of protein taken from an atomic structure of a naturally occurring protein or an ensemble of fragments of protein, the populations of which are determined either from populations in the naturally occurring proteins or from computations establishing the potential energy of each fragment and integrating them into a partition function.

36. The method of Claim 5, wherein said reference structure is a denatured state of said atomic structure and said side chain vibration entropy term in said reference structure is modeled as a uniform distribution of sub-rotamer conformations.

37. The method of Claim 18, wherein said intrinsic energy of interaction comprises at least one energy term selected from the group consisting of a molecular mechanics energy term, a solvation energy term, and an entropic contribution.

38. The method of Claim 37, wherein said intrinsic energy of interaction comprises a sum of terms whose coefficients are individually adjustable weighting factors.

39. The method of Claim 37, wherein said molecular mechanics energy term comprises at least one term selected from the group consisting of the van der Waals energy between pairs of non-bonded atoms, the hydrogen bond energy between pairs of hydrogen bond donor and acceptor atoms, and the electrostatic interaction energy between pairs of charged atoms.

40. The method of Claim 39, wherein the van der Waals energy between pairs of non-bonded atoms is calculated as

$$E_{vdw} = \sum_{\text{nonbonded } i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right)$$

wherein i and j represent all possible atom pairs comprising atoms i of said atomic structure held in a fixed geometry and atoms j of said side chain rotamer.

41. The method of Claim 39, wherein the hydrogen bonding energy is calculated as

$$E_{HB} = \sum_{\text{H-bonded } H,A} \left(\frac{A_{HA}}{r_{ij}^{12}} - \frac{B_{HA}}{r_{ij}^{10}} \right)$$

wherein the sum runs over all hydrogen bonds between said atomic structure held in a fixed geometry and said side chain rotamer, and atoms i and j are respectively donor and acceptor atoms participating in each of said hydrogen bonds.

42. The method of Claim 39, wherein the electrostatic energy between pairs of charged atoms is calculated as

$$E_{\text{elec}} = \sum_{\text{charges } ij} \frac{q_i q_j e^2}{4\pi \varepsilon_0 \varepsilon_r r_{ij}}$$

wherein the sum runs over all pairs of charged atoms such that atom i is found in said atomic structure and atom j is found in said side chain rotamer and whose respective charges are q_i and q_j.

43. The method of Claim 37, wherein said entropic contribution comprises at least one term selected from the group consisting of a main chain entropy term and a side chain vibration entropy term.

44. The method of Claim 43, wherein said main chain entropy term is calculated as

$$-w_{\text{mainchain}}^{\text{entropy}} RT_{phys} \ln \frac{\sum_{\substack{\text{subspaces } \phi\psi_{20^\circ \times 20^\circ} \\ \text{close to residue } \phi\psi}} w_{\phi\psi_{20^\circ \times 20^\circ}} N_{\phi\psi_{20^\circ \times 20^\circ}}^{\substack{\text{residue} \\ \text{type}}}}{N_{\text{all } \phi\psi}^{\substack{\text{residue} \\ \text{type}}}}$$

wherein $w_{\text{mainchain}}^{\text{entropy}}$ is a coefficient, $w_{\phi\psi}$ is obtained from the partition function, T_{phys} is physiological temperature, and N is a number of amino acids of a particular type found in a given range of main chain dihedral angles.

45. The method of Claim 43, wherein said side chain vibration entropy term is calculated as

$$-w_{\text{side chain}}^{\text{vibration}} T_{phys} \left(\left(-R \sum_{\text{sub-rotamers } s} w_s \ln w_s \right)_{\substack{\text{target} \\ \text{structure}}} - VIB_{\substack{\text{reference} \\ \text{structure}}}^{\substack{\text{residue} \\ \text{type}}} \right)$$

wherein $w_{\text{side chain}}^{\text{vibration}}$ is a coefficient, and w_s is obtained from a partition function.

46. The method of Claim 37, wherein said solvation term is calculated as

$$w^{\text{solvation}} \sum_{\text{atoms } i \text{ of side chain}} \sigma_i \left(\left(ASA_i \right)_{\substack{\text{reference} \\ \text{structure}}} - \left(ASA_i \right)_{\substack{\text{target} \\ \text{structure}}} \right)$$

wherein $w^{\text{solvation}}$ is a coefficient, and ASA_i is the accessible solvent area of atom i.

47. The method of Claim 19, wherein said pairwise energy of interaction comprises at least one energy term selected from the group consisting of a molecular mechanics term, a solvation energy term, a penalty term, and an entropic contribution.

48. The method of Claim 47, wherein said pairwise energy of interaction comprises a sum of terms whose coefficients are individually adjustable weighting factors.

49. The method of Claim 47, wherein said molecular mechanics term comprises at least one term selected from the group consisting of a van der Waals energy term, a hydrogen bond energy term and an electrostatic interaction energy term.

50. The method of Claim 49, wherein the van der Waals energy term is calculated as

$$E_{\text{vdw}} = \sum_{\text{nonbonded } ij} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right)$$

wherein the summation runs over all possible atom pairs comprising atoms i from the first side chain rotamer of one of said pairs of side chain rotamers and atoms j from the second side chain rotamer of one of said pairs of side chain rotamers.

51. The method of Claim 49, wherein the hydrogen bonding energy term is calculated as

$$E_{HB} = \Sigma_{\text{H-bonded } H,A}\left(\frac{A_{HA}}{r_{ij}^{12}} - \frac{B_{HA}}{r_{ij}^{10}}\right)$$

wherein the summation runs over all possible hydrogen bonds between atom pairs comprising atoms *i* from the first side chain rotamer of one of said pairs of side chain rotamers and atoms *j* from the second side chain rotamer of one of said pairs of side chain rotamers.

52. The method of Claim 49, wherein the electrostatic interaction energy term is calculated as

$$E_{\text{elec}} = \Sigma_{\text{charges } ij}\frac{q_i q_j e^2}{4\pi\varepsilon_0\varepsilon_r r_{ij}}$$

wherein *i* and *j* represent all possible pairs of charged atoms comprising atoms *i*, with charge q_i, from the first side chain rotamer of one of said pairs of side chain rotamers and atoms *j*, with charge q_j, from the second side chain rotamer of one of said pairs of side chain rotamers.

53. The method of Claim 47 wherein said entropic contribution comprises a side chain vibration entropy term.

54. The method of Claim 53, wherein said side chain vibration entropy term is calculated according to an energy-based distribution of sub-rotamers.

55. The method of Claim 53 wherein said side chain vibrational entropy term is calculated as:

$$-w_{\text{Pairwise}}^{\text{vibration}} T_{phys} \lambda \left(\left(-R \sum_{\substack{\text{sub-rotamers } A_i \text{ and } B_j \\ \text{of rotamer pair } AB}} w_{A_i B_j} \ln w_{A_i B_j} \right)_{\substack{\text{rotamer pair } AB \\ \text{in target structure}}} - \left(-R \sum_{\substack{\text{sub-rotamers } A_i \\ \text{of rotamer } A}} w_{A_i} \ln w_{A_i} \right)_{\substack{\text{only rotamer } A \\ \text{in target structure}}} - \left(-R \sum_{\substack{\text{sub-rotamers } B_j \\ \text{of rotamer } B}} w_{B_j} \ln w_{B_j} \right)_{\substack{\text{only rotamer } B \\ \text{in target structure}}} \right)$$

56. The method of Claim 55, wherein the weights, w_i, of the sub-rotamers in said side chain vibration entropy term are obtained by means of a partition function.

57. The method of Claim 47, wherein said solvation energy term is calculated as a difference between an accessible surface area of the side chain in the reference state and the measured accessible surface area of the side chain conformer substituted into the atomic structure and the reference structure is a denatured form of the macromolecule.

58. The method of Claim 47, wherein said solvation term is calculated as

$$w^{\text{solvation}} \sum_{\substack{\text{atoms } i \text{ of residue } A \text{ or } B \\ \text{of residue pair } AB}} \sigma_i \lambda \left(\begin{matrix} (ASA_i)_{\substack{\text{only residue } A \text{ or } B \\ \text{in target structure}}} \\ -(ASA_i)_{\substack{\text{residue pair } AB \\ \text{in target structure}}} \end{matrix} \right)$$

wherein $\left(ASA_i\right)_{\substack{\text{only residue } A \text{ or } B \\ \text{in target structure}}}$ is a measured accessible surface area of atom i in each side chain conformer substituted, separately, into the atomic structure, and $\left(ASA_i\right)_{\substack{\text{residue pair } AB \\ \text{in target structure}}}$ is the measured accessible surface area of atom i in each side chain substituted together into the target atomic structure.

59. The method of Claim 47, wherein said penalty term is calculated as

$$-w^{stat} RT_{stat} \sum_{\text{all residues } i} \ln P^{\text{stat}}_{\text{amino acid } i}$$

wherein $w^{\text{stat}}_{\text{pairwise}}$ is a coefficient, and P is a probability of occurrence of a pair of amino acid residues whose side chain rotamers are contributing to the pairwise interaction energy.

60. The method of Claim 17, wherein said intrinsic energy of interaction comprises a molecular mechanics term and said pairwise energy of interaction comprises a molecular mechanics term, and wherein the dihedral angles of said side chain rotamer are optimized in step (c) by minimizing both the molecular mechanics term of said intrinsic energy of interaction, and said molecular mechanics term of said pairwise energy of interaction.

61. The method of Claim 60, wherein the dihedral angles of said side chain rotamer are optimized in the space of internal rotations of said rotamers by an algorithm selected from the group consisting of least squares, steepest descents, quasi-Newtonian, and simulated annealing.

62. The method of Claim 60, wherein the dihedral angles of said side chain rotamer are optimized by averaging over the values measured by sampling sub-rotamer configurations of the side chain rotamers, generating said sub-rotamer configurations by sampling a range of dihedral angles by stepping each dihedral angle in said side chain

rotamer by a predetermined step size for a number of steps and, selecting said sub-rotamers whose contribution to the calculated solution score is highest.

63. The method of Claim 62, wherein the predetermined step size and the number of steps sampled is determined by an amino acid type of the side chain rotamer.

64. The method of Claim 6, wherein the side chain rotamers which are not rejected and that have a probability smaller than a predetermined probability threshold are rejected.

65. The method of Claim 64, wherein said probability is calculated from a partition function over all possible rotamers of a particular residue type, using their energy of interaction with a fixed portion of the atomic structure wherein said energy of interaction comprises at least one energy term selected from the group consisting of a molecular mechanics energy term, a solvation energy term, and an entropic contribution.

66. The method of Claim 2, wherein each candidate conformer specified in step (b) is a D-enantiomer selected from the group consisting of glycine, alanine, valine, leucine, isoleucine, glutamic acid, aspartic acid, asparagine, glutamine, proline, phenylalanine, tyrosine, serine, threonine, lysine, arginine, histidine, cysteine, tryptophan, and methionine.

67. The method of Claim 2, wherein each candidate conformer specified in step (b) is an L-enantiomer selected from the group consisting of: glycine, alanine, valine, leucine, isoleucine, glutamic acid, aspartic acid, asparagine, glutamine, proline, phenylalanine, tyrosine, serine, threonine, lysine, arginine, histidine, cysteine, tryptophan, and methionine.

68. The method of Claim 2, wherein each candidate conformer specified in step (b) is selected from the group consisting of the L- and D-enantiomers of amino acids including, but not limited to, norvaline, beta-alanine, and tartrine.

69. The method of Claim 11, wherein said library of predetermined rotamer conformations is constructed by a method comprising:

tabulating, for each of the twenty naturally occurring amino acids, a statistical distribution of observed amino acid side chain dihedral angles in a set of crystallographically determined protein structures;

determining a Gaussian distribution of observed amino acid side chain dihedral angles tabulated in said tabulating step; and

constructing amino acid side chain rotamers for each of the twenty naturally occurring amino acids using all combinations of Gaussian peaks around each amino acid side chain dihedral angle tabulated in said tabulating step.

70. The method of Claim 11, wherein said library of predetermined rotamer conformations is constructed *ab initio*, by a method comprising: computing portions of vibration-rotation potential energy surfaces of said side chain rotamers and determining, through exhaustive sampling, dihedral angles at which minima are found on said vibration-rotation potential energy surfaces.

71. The method of Claim 2, further comprising the step of storing all side chain rotamers having a calculated solution score that is lower than said threshold value in an array and eliminating a subset of side chain rotamers from said array using dead-end elimination where energy terms are partitioned according to:

$$Template\ Energy + \sum_{\substack{residues \\ i}} Intrinsic\ Energy + \sum_{\substack{residues \\ i}} \sum_{\substack{residues \\ j>i}} Pairwise\ Energy$$

72. The method of Claim 71, wherein dead-end elimination comprises elimination of a side chain rotamer *r*, of residue *i*, when the inequality

$$E_{i_r}^{template} - E_{i_t}^{template} + \sum_{j \neq i} \min_s \left(E_{i_r j_s}^{pairwise} - E_{i_t j_s}^{pairwise} \right) > 0$$

is true.

73. The method of Claim 2, further comprising the step of eliminating a side chain rotamer pair using dead-end elimination, wherein a side chain rotamer pair comprises a first side chain rotamer corresponding to a first building block in said specified set of substitute building blocks and a second side chain rotamer corresponding to a second building block in said specified set of substitute building blocks.

74. The method of Claim 73, wherein dead-end elimination comprises eliminating a side chain rotamer pair that consists of rotamer r, of residue i, and rotamer s, of residue j, when the inequalities:

$$E_{(i_r,j_s)} + \sum_{k \neq i,j} \min_t\left(E_{(i_r,j_s),k_t}\right) > E_{(i_u,j_v)} + \sum_{k \neq i,j} \max_t\left(E_{(i_u,j_v),k_t}\right)$$

and

$$E_{(i_r,j_s)} - E_{(i_u,j_v)} + \sum_{k \neq i,j} \min_t\left(E_{(i_r,j_s),k_t} - E_{(i_u,j_v),k_t}\right) > 0$$

are true.

75. The method of Claim 71, wherein the eliminating step is repeated until no additional side chain rotamers are eliminated from said array by the process of dead-end elimination.

76. The method of Claim 73, wherein the eliminating step is repeated until no additional side chain rotamer pairs are eliminated from said array by the process of dead-end elimination.

77. The method of Claim 2, wherein the calculated solution score additionally comprises an entropy term that is determined by a difference between an entropy of a side chain rotamer in the solution when substituted into the atomic structure, and an entropy of the side chain rotamer in the solution when in a denatured state.

78. The method of Claim 77 wherein entropy contributions of said side chain rotamers are derived from experimental or empirical data.

79. The method of Claim 77, wherein the entropy term is computed using an iterative

method.

80. The method of Claim 79, wherein the entropy term is computed using mean field theory.

81. The method of Claim 80, wherein the contribution of the rotamers to the entropy may be split into intrinsic and pairwise terms.

82. The method of Claim 2 wherein said calculated solution score comprises a difference between a first value corresponding to said solution structure and a second value corresponding to a weighted average over a plurality of reference structures.

83. The method of Claim 1 wherein said target macromolecule consists of a plurality of structures at least one of which is represented by an atomic structure.

84. A computer program product for use in conjunction with a computer, the computer program product comprising a computer readable storage medium and a computer

program mechanism embedded therein, the computer program mechanism comprising an optimizer module configured to choose a set of substitute building blocks for a set of positions in a target macromolecule according to whether a calculated solution score is lower than a threshold value, the computer program mechanism, upon receiving as input said set of positions:

(a) specifying at least one substitute building block for each position in said set of positions to produce a specified set of substitute building blocks;

(b) for each said substitute building block

i) determining at least one candidate conformer;

ii) substituting coordinates of each said candidate conformer or portion thereof for coordinates of the building block or portion thereof at said position in an atomic structure of said target macromolecule;

(c) minimizing the value of a calculated energy term by adjusting the geometry of each said candidate conformer or portion thereof in order to obtain a solution structure;

(d) calculating a solution score for said solution structure, wherein said solution score comprises an entropic term; and

(e) choosing said specified set of substitute building blocks if said calculated solution score is lower than a threshold value.

85. The computer program product of Claim 84, wherein said macromolecule is a peptide or protein; said building blocks are amino acid residues; and each candidate conformer is a side chain rotamer selected from a plurality of side chain rotamers.

86. The computer program product of Claim 85 wherein said calculated solution score comprises a difference between a first value corresponding to said solution structure and a second value corresponding to a reference structure

87. The computer program product of Claim 86, wherein said first value corresponding to said solution structure accounts for interactions between said side chain rotamer and said atomic structure, and a sum of interactions between all pairs of all possible side chain rotamers.

88. The computer program product of Claim 87, further comprising a step of rejecting a side chain rotamer when the value of said interactions between said side chain rotamer and said atomic structure is greater than a threshold value.

89. The computer program product of Claim 86, wherein said reference structure is a denatured state of said solution structure.

90. The computer program product of Claim 85, wherein the dihedral angles of said side chain rotamer are optimized in step (c).

91. The computer program product of Claim 85, wherein the positions of all main chain atoms of said atomic structure, and the positions of all atoms in amino acid side chains that are not included in said set of substitute building blocks are held fixed in said atomic structure.

92. The computer program product of Claim 90, wherein the positions of all atoms in amino acid side chains on residues that are not at said set of positions are allowed to vary whilst the dihedral angles of said side chain rotamer are optimized.

93. The computer program product of Claim 90, wherein the positions of all main chain atoms of said atomic structure are allowed to vary whilst the dihedral angles of said side chain rotamer are optimized.

94. The computer program product of Claim 85, wherein said plurality of conformers is a library of predetermined side chain rotamer conformations.

95. The computer program product of Claim 85, wherein at least one side chain conformer in said plurality of side chain rotamers is derived from a continuous distribution of conformations.

96. The computer program product of Claim 84, wherein said atomic structure includes a representation of the building blocks at each position in said set of positions; and said atomic structure was determined by a method selected from the group consisting of x-ray crystallography, nuclear magnetic resonance spectroscopy, electron microscopy, homology modeling, and *ab initio* molecular modeling.

97. The computer program product of Claim 84, wherein said atomic structure is an X-ray crystal structure of a portion of said macromolecule that comprises said building blocks at each position.

98. The computer program product of Claim 97, wherein said X-ray crystal structure was determined at a resolution of less than 4.0 Angstroms.

99. The computer program product of Claim 85, wherein said solution score is calculated using an empirical scoring function.

100. The computer program product of Claim 99, wherein said empirical scoring function is a sum of energy terms, comprising a template energy of said atomic structure held in a fixed geometry, an intrinsic energy of interaction between a side chain rotamer and said atomic structure held in a fixed geometry and a pairwise energy of interaction between possible pairs of side chain rotamers in said substitute set of building blocks.

101. The computer program product of Claim 100, wherein the template energy of said atomic structure comprises at least one energy term selected from the group consisting of a molecular mechanics potential, a solvation energy, an empirical penalty function, and an entropic contribution.

102. The computer program product of Claim 101, wherein the template energy of said atomic structure comprises a sum of terms whose coefficients are individually adjustable weighting factors.

103. The computer program product of Claim 102, wherein said molecular mechanics potential comprises at least one energy term selected from the group consisting of bond length vibrations, bond angle bends, the hydrogen bond energy between pairs of hydrogen bond donor and acceptor atoms, an electrostatic interaction energy between pairs of charged atoms, and a van der Waals interaction energy between pairs of non-bonded atoms in said atomic structure.

104. The computer program product of Claim 101, wherein said entropic contribution comprises at least one term selected from the group consisting of a main chain entropy term, a side chain rotation entropy term and a side chain vibration entropy term.

105. The computer program product of Claim 89, wherein said reference structure comprises said side chain rotamer embedded in an alanine based penta-peptide.

106. The computer program product of Claim 89, wherein said reference structure comprises said side chain rotamer embedded in a fragment of protein taken from an atomic structure of a naturally occurring protein or an ensemble of fragments of protein, the populations of which are determined either from populations in the naturally occurring proteins or from computations establishing the potential energy of each fragment and integrating them into a partition function.

107. The computer program product of Claim 88, wherein the side chain rotamers which are not rejected and that have a probability smaller than a predetermined probability threshold are rejected.

108. The computer program product of Claim 94, wherein said library of predetermined rotamer conformations is constructed by a method comprising:

tabulating, for each of the twenty naturally occurring amino acids, a statistical distribution of observed amino acid side chain dihedral angles in a set of crystallographically determined protein structures;

determining a Gaussian distribution of observed amino acid side chain dihedral angles tabulated in said tabulating step; and

constructing amino acid side chain rotamers for each of the twenty naturally occurring amino acids using all combinations of Gaussian peaks around each amino acid side chain dihedral angle tabulated in said tabulating step.

109. The computer program product of Claim 94, wherein said library of predetermined rotamer conformations is constructed *ab initio*, by a method comprising: computing portions of vibration-rotation potential energy surfaces of said side chain rotamers and determining, through exhaustive sampling, dihedral angles at which minima are found on said vibration-rotation potential energy surfaces.

110. The computer program product of Claim 85, further comprising the step of storing all side chain rotamers having a solution score that is lower than said threshold value in an array and eliminating a subset of side chain rotamers from said array using dead-end elimination where energy terms are partitioned according to:

$$Template\ Energy + \sum_{\substack{residues \\ i}} Intrinsic\ Energy + \sum_{\substack{residues \\ i}} \sum_{\substack{residues \\ j>i}} Pairwise\ Energy$$

111. The computer program product of Claim 110, wherein dead-end elimination comprises elimination of a side chain rotamer r, of residue i, when the inequality

$$E_{i_r}^{\text{template}} - E_{i_t}^{\text{template}} + \sum_{j \neq i} \min_s \left(E_{i_r j_s}^{\text{pairwise}} - E_{i_t j_s}^{\text{pairwise}} \right) > 0$$

is true.

112. The computer program product of Claim 85, further comprising the step of eliminating a side chain rotamer pair using dead-end elimination, wherein a side chain rotamer pair comprises a first side chain rotamer representing a first building block in said set of building blocks and a second side chain rotamer representing a second building block in said set of building blocks.

113. The computer program product of Claim 112, wherein dead-end elimination comprises eliminating a side chain rotamer pair that consists of rotamer r, of residue i, and rotamer s, of residue j, when the inequalities:

$$E_{(i_r, j_s)} + \sum_{k \neq i,j} \min_t \left(E_{(i_r, j_s), k_t} \right) > E_{(i_u, j_v)} + \sum_{k \neq i,j} \max_t \left(E_{(i_u, j_v), k_t} \right)$$

and

$$E_{(i_r, j_s)} - E_{(i_u, j_v)} + \sum_{k \neq i,j} \min_t \left(E_{(i_r, j_s), k_t} - E_{(i_u, j_v), k_t} \right) > 0$$

are true.

114. The computer program product of Claim 110, wherein the eliminating step is repeated until no additional side chain rotamers are eliminated from said array by the process of dead-end elimination.

115. The computer program product of Claim 112, wherein the eliminating step is repeated until no additional side chain rotamer pairs are eliminated from said array by the process of dead-end elimination.

116. The computer program product of Claim 85, wherein the solution score additionally comprises an entropy term that is determined by a difference between an entropy of a side chain rotamer in the solution when substituted into the atomic structure, and an entropy of the side chain rotamer in the solution when in a denatured state.

117. The computer program product of Claim 116, wherein entropy contributions of said side chain rotamers are derived from experimental or empirical data.

118. The computer program product of Claim 116, wherein the entropy term is computed using an iterative method.

119. The computer program product of Claim 118, wherein the entropy term is computed using mean field theory.

120. The computer program product of Claim 119, wherein the contribution of the rotamers to the entropy may be split into intrinsic and pairwise terms.

121. A system for choosing a set of substitute building blocks for a set of positions in a target macromolecule according to whether a calculated solution score is lower than a threshold value, comprising:

a central processing unit;

an input device for inputting requests;

an output device;

a memory;

at least one bus connected to the central processing unit, the memory, the input device, and the output device;

the memory storing an computer program mechanism comprising an optimizer module configured to choose the set of substitute building blocks, the computer program mechanism, upon receiving a request to choose the set of substitute building blocks,

(a) specifying at least one substitute building block for each position in said set of positions to produce a specified set of substitute building blocks;

(b) for each said substitute building block,

i) determining at least one candidate conformer;

ii) substituting coordinates of each said candidate conformer or portion thereof for coordinates of the building block or portion thereof at said position in an atomic structure of said target macromolecule;

(c) minimizing the value of a calculated energy term by adjusting the geometry of each said candidate conformer or portion thereof in order to obtain a solution structure;

(d) calculating a solution score for said solution structure, wherein said solution score comprises an entropic term; and

(e) choosing said specified set of substitute building blocks if said calculated solution score is lower than a threshold value.

122. The system of Claim 121, wherein said macromolecule is a peptide or protein; said building blocks are amino acid residues; and each candidate conformer is a side chain rotamer selected from a plurality of side chain rotamers.

123. The system of Claim 122 wherein said calculated solution score comprises a difference between a first value corresponding to said solution structure and a second value corresponding to a reference structure.

124. The system of Claim 123, wherein said first value corresponding to said solution structure accounts for interactions between said side chain rotamer and said atomic structure, and a sum of interactions between all pairs of all possible side chain rotamers.

125. The system of Claim 124, further comprising a step of rejecting a side chain rotamer when the value of said interactions between said side chain rotamer and said atomic structure is greater than a threshold value.

126. The system of Claim 123, wherein said reference structure is a denatured state of said solution structure.

127. The system of Claim 122, additionally comprising a step wherein the dihedral angles of said side chain rotamer are optimized in step (c).

128. The system of Claim 122, wherein the positions of all main chain atoms of said atomic structure, and the positions of all atoms in amino acid side chains that are not included in said set of substitute building blocks are held fixed in said atomic structure.

129. The system of Claim 127, wherein the positions of all atoms in amino acid side chains on residues that are not at said set of positions are allowed to vary whilst the dihedral angles of said rotamer are optimized.

130. The system of Claim 127, wherein the positions of all main chain atoms of said atomic structure are allowed to vary whilst the dihedral angles of said rotamer are optimized.

131. The system of Claim 122, wherein said plurality of conformers is a library of predetermined rotamer conformations.

132. The system of Claim 122, wherein at least one side chain rotamer in said plurality of side chain rotamers is derived from a continuous distribution of conformations.

133. The system of Claim 121, wherein said atomic structure includes a representation of the building blocks at each position in said set of positions; and said atomic structure was determined by a method selected from the group consisting of x-ray crystallography, nuclear magnetic resonance spectroscopy, electron microscopy, homology modeling, and *ab initio* molecular modeling.

134. The system of Claim 121, wherein said atomic structure is an X-ray crystal structure of a portion of said macromolecule that comprises said building blocks at each position.

135. The system of Claim 124, wherein said X-ray crystal structure was determined at a resolution of less than 4.0 Angstroms.

136. The system of Claim 122, wherein said calculated solution score is obtained using an empirical scoring function.

137. The system of Claim 136, wherein said empirical scoring function is a sum of energy terms, comprising a template energy of said atomic structure held in a fixed

geometry, an intrinsic energy of interaction between a side chain rotamer and said atomic structure held in a fixed geometry and a pairwise energy of interaction between possible pairs of side chain rotamers in said substitute set of building blocks.

138. The system of Claim 137, wherein the template energy of said atomic structure comprises at least one energy term selected from the group consisting of a molecular mechanics potential, a solvation energy, an empirical penalty function, and an entropic contribution.

139. The system of Claim 138, wherein the template energy of said atomic structure comprises a sum of terms whose coefficients are individually adjustable weighting factors.

140. The system of Claim 139, wherein said molecular mechanics potential comprises at least one energy term selected from the group consisting of bond length vibrations, bond angle bends, the hydrogen bond energy between pairs of hydrogen bond donor and acceptor atoms, an electrostatic interaction energy between pairs of charged atoms, and a van der Waals interaction energy between pairs of non-bonded atoms in said atomic structure.

141. The system of Claim 138, wherein said entropic contribution comprises at least one term selected from the group consisting of a main chain entropy term, a side chain rotation entropy term and a side chain vibration entropy term.

142. The system of Claim 125, wherein said reference structure comprises said side chain rotamer substituted for a side chain rotamer in an alanine based penta-peptide.

143. The system of Claim 125, wherein said reference structure comprises said side chain rotamer embedded in a fragment of protein taken from an atomic structure of a naturally occurring protein or an ensemble of fragments of protein, the populations of which

are determined either from populations in the naturally occurring proteins or from computations establishing the potential energy of each fragment and integrating them into a partition function.

144. The system of Claim 124, wherein the side chain rotamers which are not rejected in and that have a probability smaller than a predetermined probability threshold are rejected.

145. The system of Claim 131, wherein said library of predetermined rotamer conformations is constructed by a method comprising:

tabulating, for each of the twenty naturally occurring amino acids, a statistical distribution of observed amino acid side chain dihedral angles in a set of crystallographically determined protein structures;

determining a Gaussian distribution of observed amino acid side chain dihedral angles tabulated in said tabulating step; and

constructing amino acid side chain rotamers for each of the twenty naturally occurring amino acids using all combinations of Gaussian peaks around each amino acid side chain dihedral angle tabulated in said tabulating step.

146. The system of Claim 131, wherein said library of predetermined rotamer conformations is constructed *ab initio*, by a method comprising: computing portions of vibration-rotation potential energy surfaces of said side chain rotamers and determining, through exhaustive sampling, dihedral angles at which minima are found on said vibration-rotation potential energy surfaces.

147. The system of Claim 122, further comprising the step of storing all side chain rotamers having a solution score that is lower than said threshold value in an array and eliminating a subset of side chain rotamers from said array using dead-end elimination where energy terms are partitioned according to:

$$Template\ Energy + \sum_{\substack{residues \\ i}} Intrinsic\ Energy + \sum_{\substack{residues \\ i}} \sum_{\substack{residues \\ j>i}} Pairwise\ Energy$$

148. The system of Claim 147, wherein dead-end elimination comprises elimination of a side chain rotamer r, of residue i, when the inequality

$$E_{l_r}^{template} - E_{l_r}^{template} + \sum_{j \neq i} \min_s \left(E_{i_r, j_s}^{pairwise} - E_{l_r, j_s}^{pairwise} \right) > 0$$

is true.

149. The system of Claim 122, further comprising the step of eliminating a side chain rotamer pair using dead-end elimination, wherein a side chain rotamer pair comprises a first side chain rotamer representing a first building block in said specified set of substitute building blocks and a second side chain rotamer representing a second building block in said specified set of substitute building blocks.

150. The system of Claim 149, wherein dead-end elimination comprises eliminating a side chain rotamer pair that consists of rotamer r, of residue i, and rotamer s, of residue j, when the inequalities:

$$E_{(i_r, j_s)} + \sum_{k \neq i, j} \min_t \left(E_{(i_r, j_s), k_t} \right) > E_{(i_u, j_v)} + \sum_{k \neq i, j} \max_t \left(E_{(i_u, j_v), k_t} \right)$$

and

$$E_{(i_r, j_s)} - E_{(i_u, j_v)} + \sum_{k \neq i, j} \min_t \left(E_{(i_r, j_s), k_t} - E_{(i_u, j_v), k_t} \right) > 0$$

are true.

151. The system of Claim 147, wherein the eliminating step is repeated until no additional side chain rotamers are eliminated from said array by the process of dead-end elimination.

152. The system of Claim 149, wherein the eliminating step is repeated until no additional side chain rotamer pairs are eliminated from said array by the process of dead-end elimination.

153. The system of Claim 122, wherein the solution score additionally comprises an entropy term that is determined by a difference between an entropy of a side chain rotamer in the solution when substituted into the atomic structure, and an entropy of the side chain rotamer in the solution when in a denatured state.

154. The system of Claim 153, wherein entropy contributions of said side chain rotamers are derived from experimental or empirical data.

155. The system of Claim 153, wherein the entropy term is computed using an iterative method.

156. The system of Claim 153, wherein the entropy term is computed using mean field theory.

157. The system of Claim 154, wherein the contribution of the rotamers to the entropy may be split into intrinsic and pairwise terms.

158. The method of Claim 1 additionally comprising the step of synthesizing at least one of said solution structures which has a calculated solution score lower than said threshold value.

159. The method of Claim 158 additionally comprising the step of screening each of said solution structures that has been synthesized against an assay to test for activity.

160. The method of Claim 1 wherein at least one of said building blocks in said set of building blocks contacts a binding partner of said target macromolecule when bound to said target macromolecule.

161. The method of Claim 2 wherein said calculated solution score comprises a difference between a first value corresponding to said solution structure and a second value corresponding to a denatured state of said solution structure,

wherein said calculated solution score is calculated using an empirical scoring function that comprises a sum of energy terms, including an energy of said atomic structure held in a fixed geometry, an intrinsic energy of interaction between a candidate side chain rotamer and said atomic structure held in a fixed geometry and a pairwise energy of interaction between possible pairs of side chain rotamers in said substitute set of building blocks,

wherein any one of said energy of said atomic structure, said intrinsic energy of interaction and said pairwise energy of interaction comprises at least one energy term selected from the group consisting of a molecular mechanics potential, a solvation energy, an empirical penalty function, and an entropic contribution, and

wherein said entropic contribution is computed using mean field theory.

SEQUENCE LISTING

<110> The European Molecular Biology Laboratory (EMBL)

<120> A COMPUTER-BASED METHOD FOR
MACROMOLECULAR ENGINEERING AND DESIGN

<130> 9882-017-228

<150> US 09/387,741
<151> 1999-08-31

<160> 20

<170> FastSEQ for Windows Version 4.0

<210> 1
<211> 6
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 1
Ala Val Ile Leu Phe Trp
1 5

<210> 2
<211> 14
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 2
Ala Val Ile Leu Gly Asp Asn Ser Thr Glu Lys Arg Tyr Trp
1 5 10

<210> 3
<211> 4
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 3
Ala Val Ile Leu
1

<210> 4

<211> 5
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 4
Ala Val Ile Leu Phe
1 5

<210> 5
<211> 4
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 5
Ala Val Leu Phe
1

<210> 6
<211> 8
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 6
Val Ala Val Met Leu Leu Trp Val
1 5

<210> 7
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 7
Ile Val Ile Ile Leu Leu Val Ile Val
1 5

<210> 8
<211> 4
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 8
Val Gly Ser Lys
1

<210> 9
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 9
Leu Val Ile Val Leu Leu Val Ile Val
1 5

<210> 10
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 10
Val Val Ile Ile Leu Leu Val Ile Val
1 5

<210> 11
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 11
Ile Val Ile Ile Leu Leu Val Val Val
1 5

<210> 12
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 12
Leu Ile Ile Val Leu Leu Val Ile Val
1 5

<210> 13

<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 13
Ile Val Val Ile Leu Leu Val Ile Val
1 5

<210> 14
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 14
Ile Ile Ile Val Leu Leu Val Ile Val
1 5

<210> 15
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 15
Ile Val Leu Ile Leu Leu Val Ile Val
1 5

<210> 16
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 16
Leu Val Ile Ile Leu Leu Val Ile Val
1 5

<210> 17
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 17
Ile Val Ile Ile Leu Leu Val Ile Val
1 5

<210> 18
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 18
Val Ala Val Met Leu Leu Val Val Val
1 5

<210> 19
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 19
Val Val Leu Ile Leu Leu Val Ile Leu
1 5

<210> 20
<211> 9
<212> PRT
<213> Artificial Sequence

<220>
<223> Conceptual sequence for computer modeling

<400> 20
Val Val Leu Leu Leu Leu Ala Phe Leu
1 5

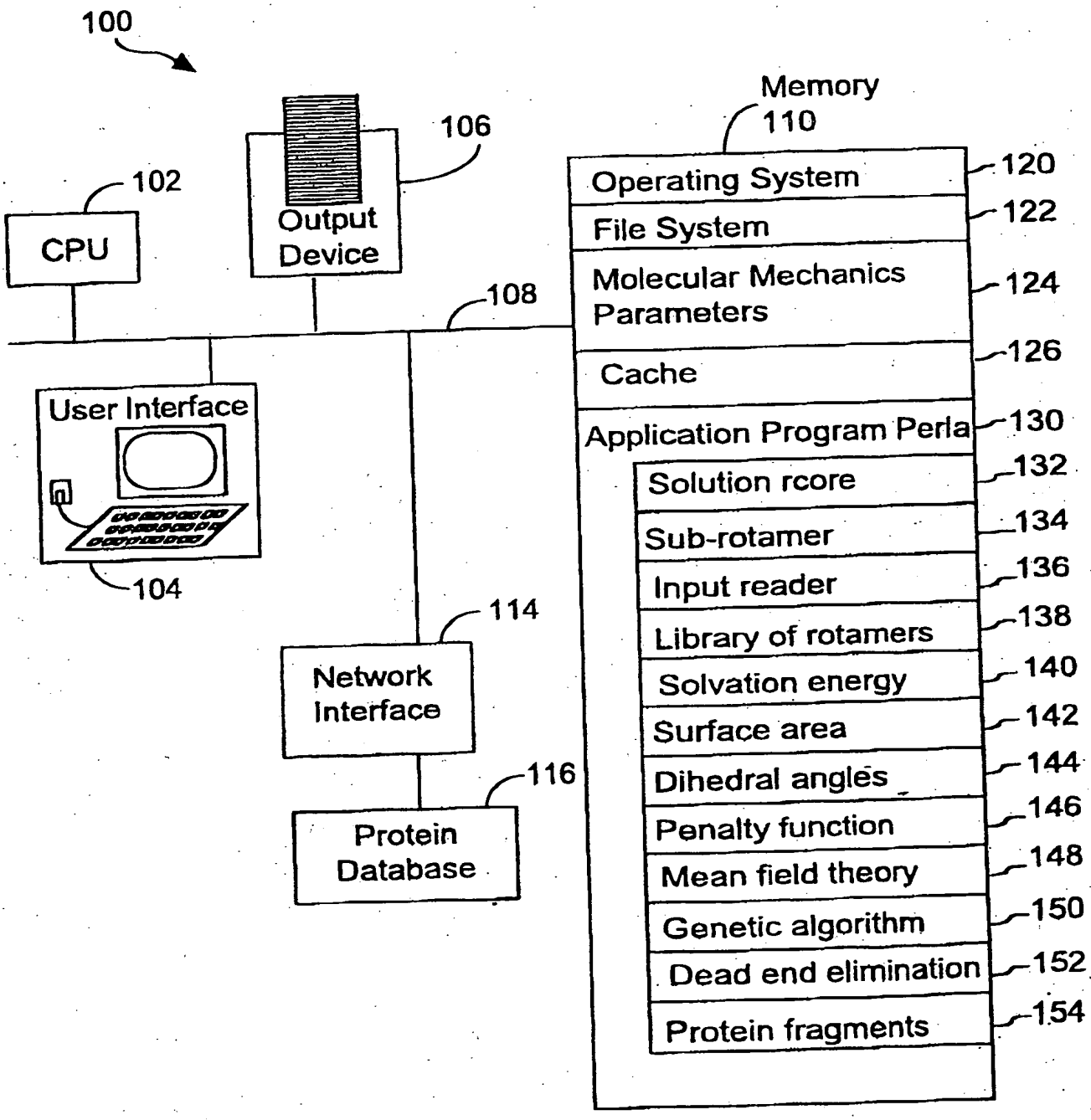

100
Memory
110
106
102
CPU
Output Device
108
Operating System
120
File System
122
Molecular Mechanics Parameters
124
Cache
126
User Interface
Application Program Perla
130
Solution rcore
132
Sub-rotamer
134
Input reader
136
Library of rotamers
138
Solvation energy
140
Surface area
142
Dihedral angles
144
Penalty function
146
Mean field theory
148
Genetic algorithm
150
Dead end elimination
152
Protein fragments
154
104
114
Network Interface
116
Protein Database

FIG. 1

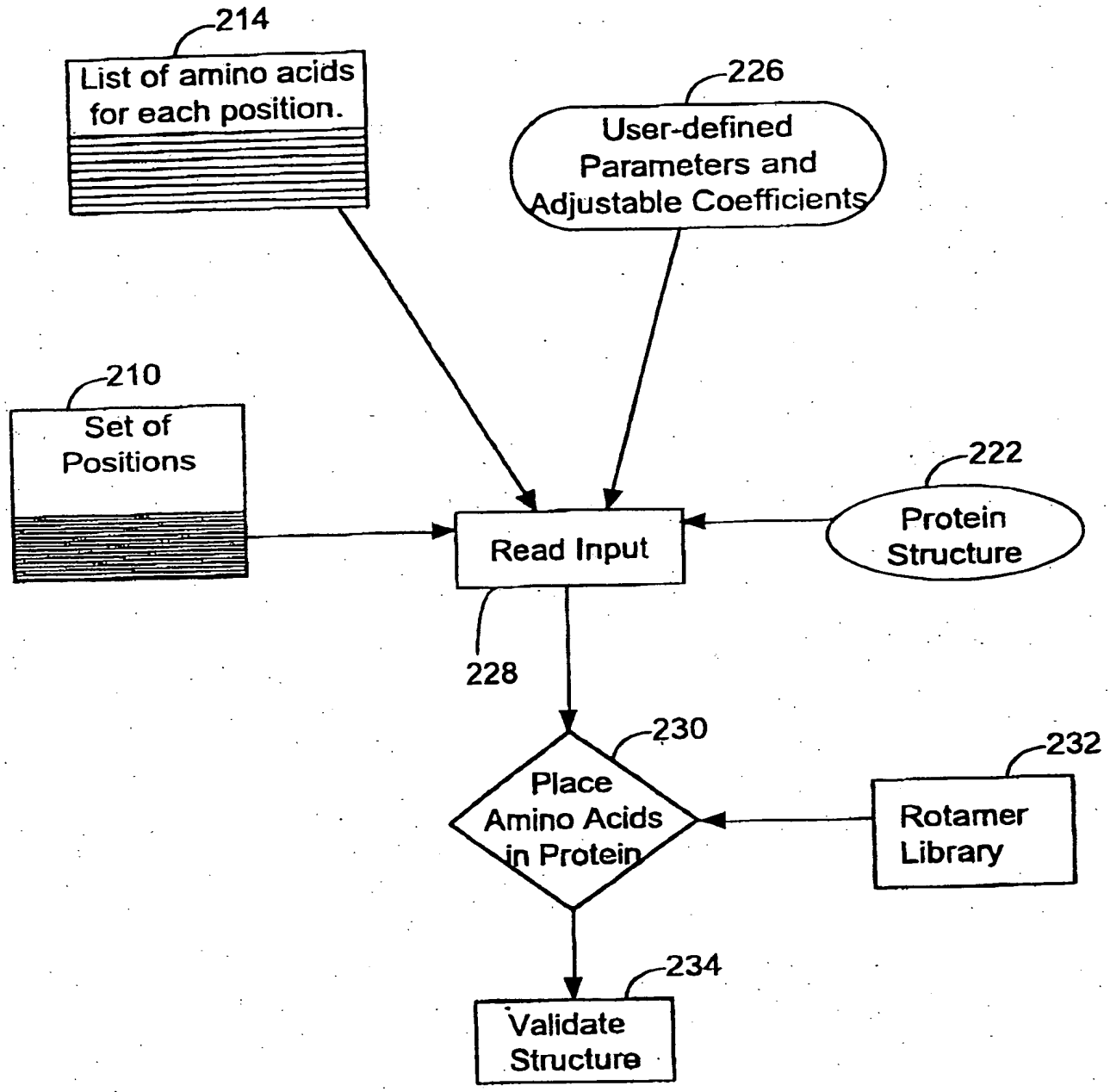

214
List of amino acids for each position.
226
User-defined Parameters and Adjustable Coefficients
210
Set of Positions
222
Protein Structure
Read Input
228
230
Place Amino Acids in Protein
232
Rotamer Library
234
Validate Structure

FIG. 2

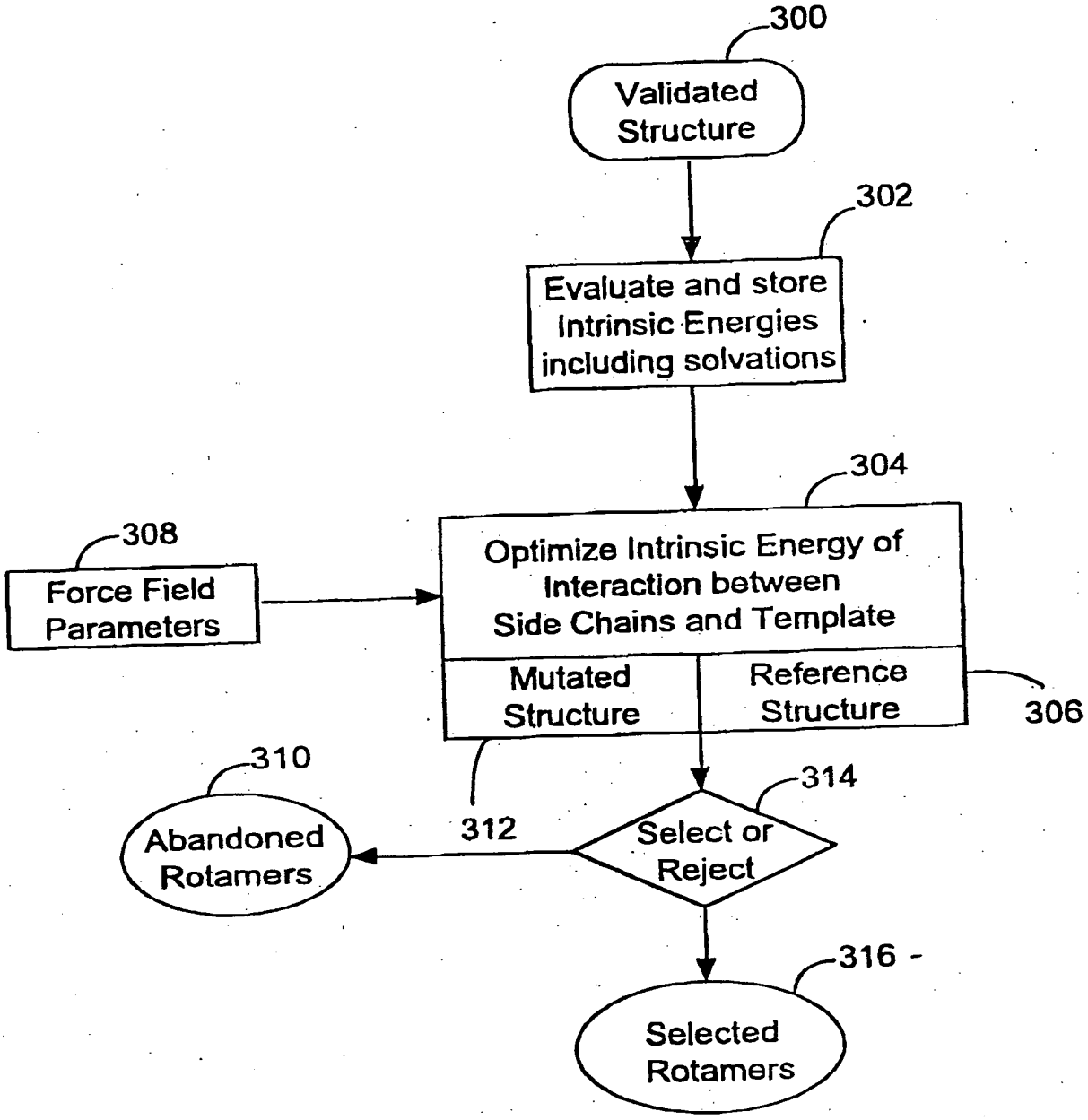

300
Validated Structure
302
Evaluate and store Intrinsic Energies including solvations
304
Optimize Intrinsic Energy of Interaction between Side Chains and Template
Mutated Structure
Reference Structure
306
308
Force Field Parameters
310
Abandoned Rotamers
312
314
Select or Reject
316
Selected Rotamers

FIG. 3

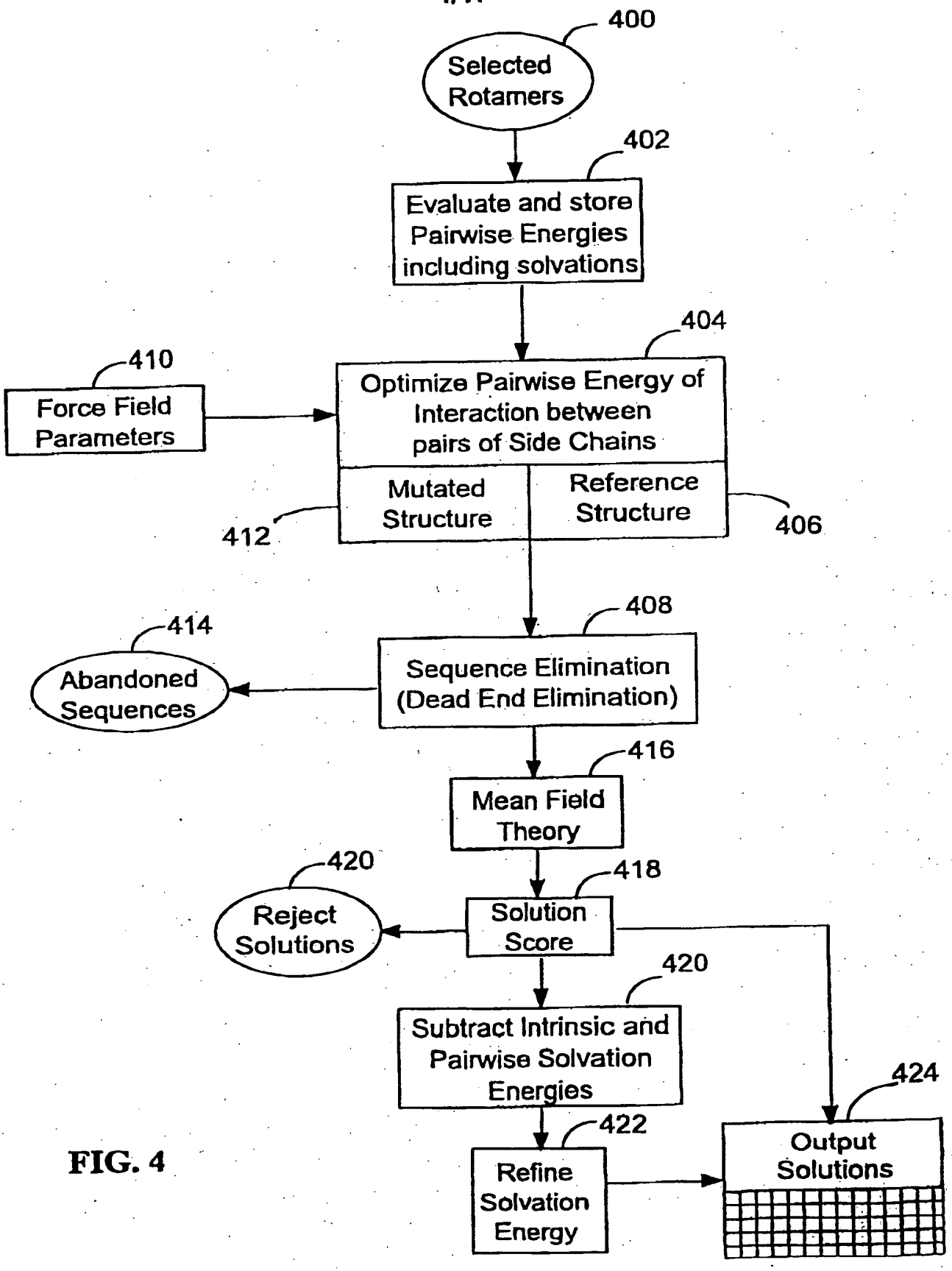

400
Selected Rotamers
402
Evaluate and store Pairwise Energies including solvations
404
410
Force Field Parameters
Optimize Pairwise Energy of Interaction between pairs of Side Chains
Mutated Structure
Reference Structure
412
406
408
414
Abandoned Sequences
Sequence Elimination (Dead End Elimination)
416
Mean Field Theory
420
418
Reject Solutions
Solution Score
420
Subtract Intrinsic and Pairwise Solvation Energies
424
422
Refine Solvation Energy
Output Solutions

FIG. 4

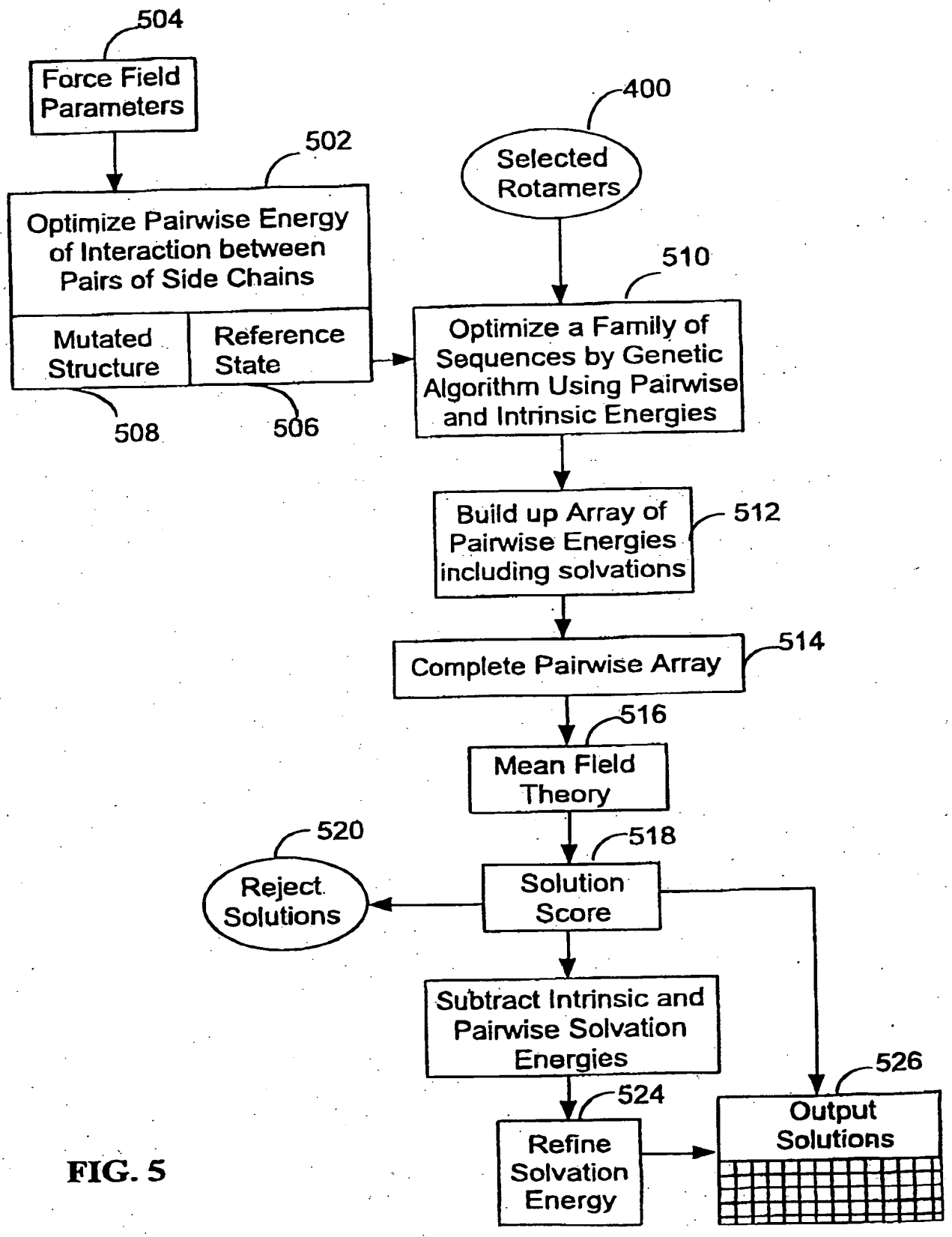

504
Force Field Parameters
502
Optimize Pairwise Energy of Interaction between Pairs of Side Chains
Mutated Structure
Reference State
508
506
400
Selected Rotamers
510
Optimize a Family of Sequences by Genetic Algorithm Using Pairwise and Intrinsic Energies
Build up Array of Pairwise Energies including solvations
512
Complete Pairwise Array
514
516
Mean Field Theory
520
Reject Solutions
518
Solution Score
Subtract Intrinsic and Pairwise Solvation Energies
524
Refine Solvation Energy
526
Output Solutions

FIG. 5

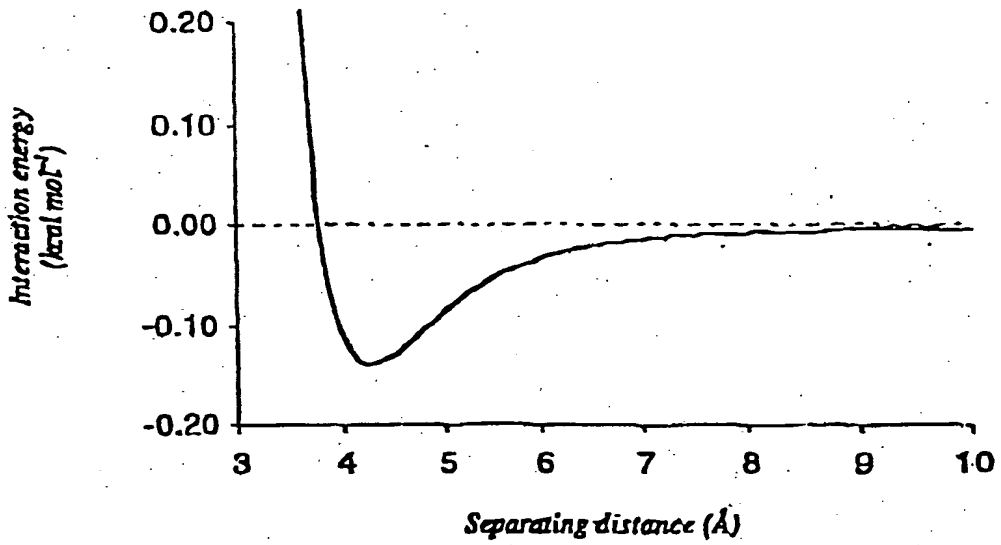

0.20
0.10
0.00
-0.10
-0.20
3
4
5
6
7
8
9
10
Interaction energy (kcal mol⁻¹)
Separating distance (Å)

FIG. 6

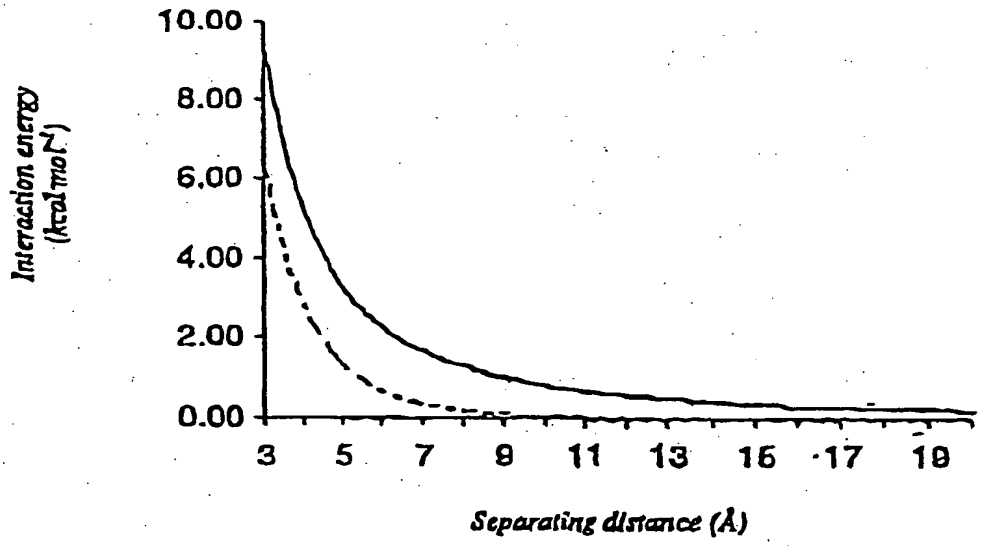
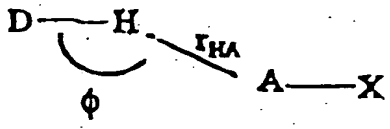

10.00
8.00
6.00
4.00
2.00
0.00
3
5
7
9
11
13
15
17
19
Interaction energy (kcal mol⁻¹)
Separating distance (Å)

FIG. 7



$1.7\text{Å} \leq r_{HA} \leq 2.5\text{Å}$

$\phi > 100°$

FIG. 8

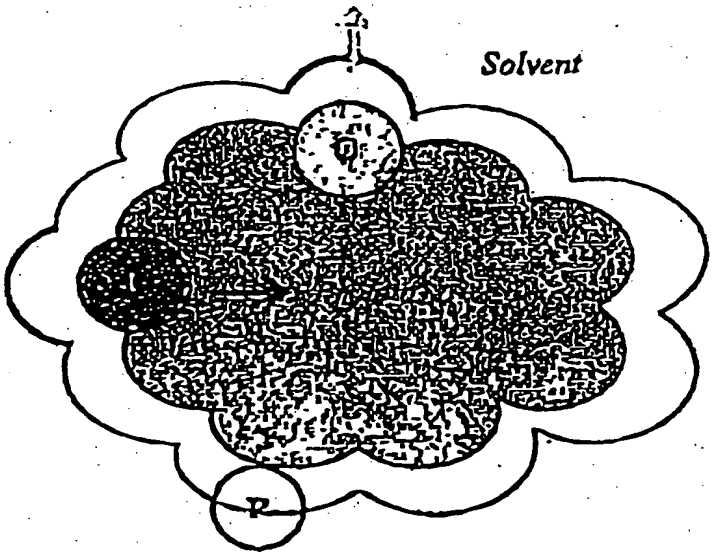


FIG. 9

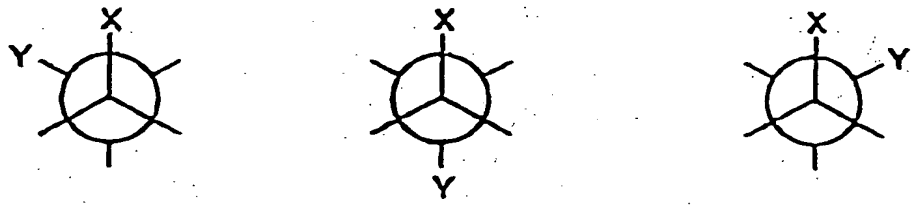


FIG. 10

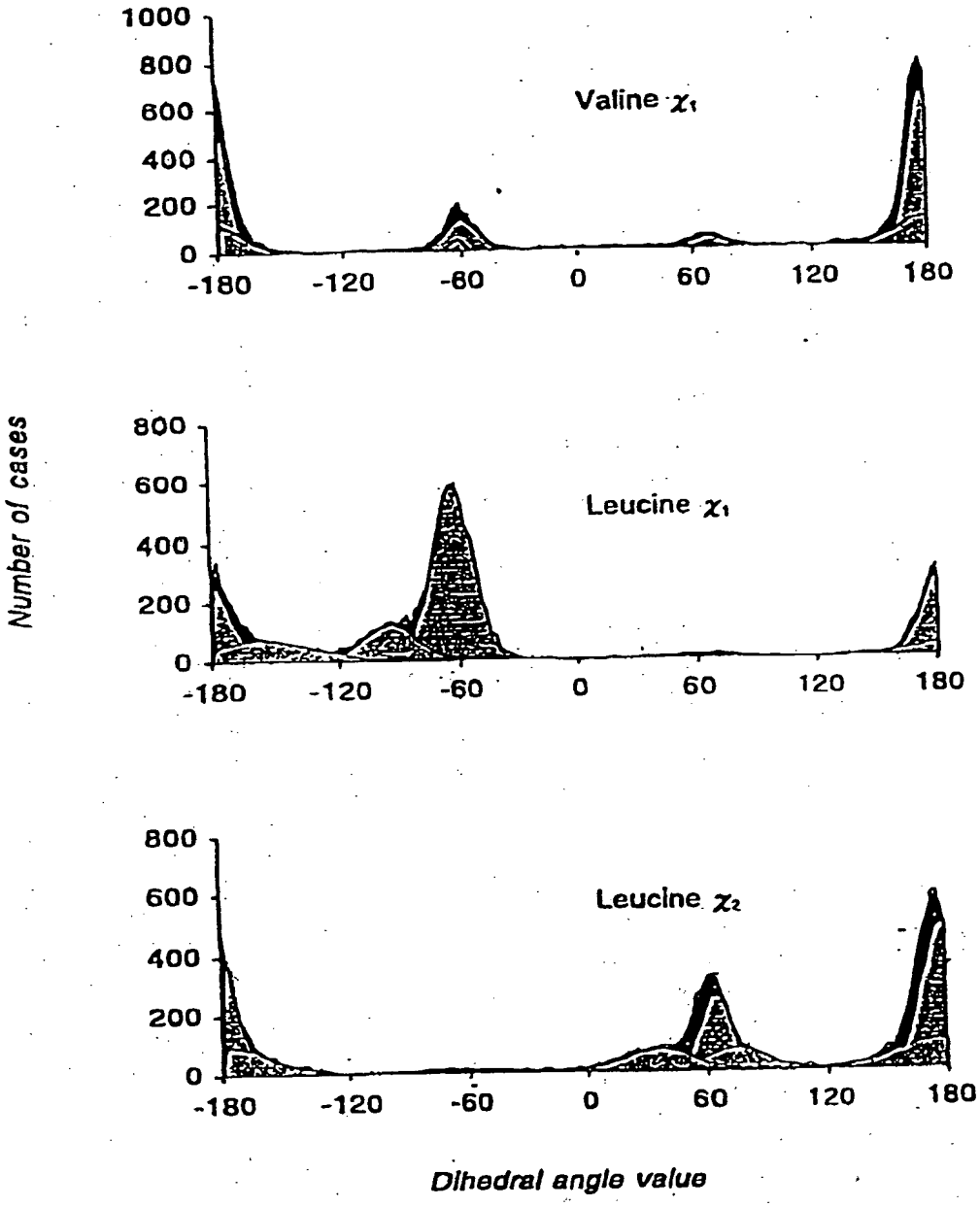
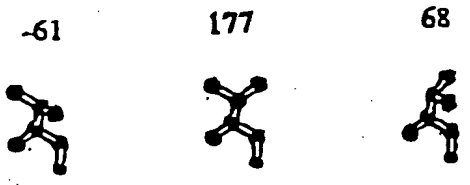

Valine χ_1
Leucine χ_1
Leucine χ_2
Number of cases
Dihedral angle value
1000
800
600
400
200
0
-180
-120
-60
0
60
120
180

FIG. 11

Valine



Leucine

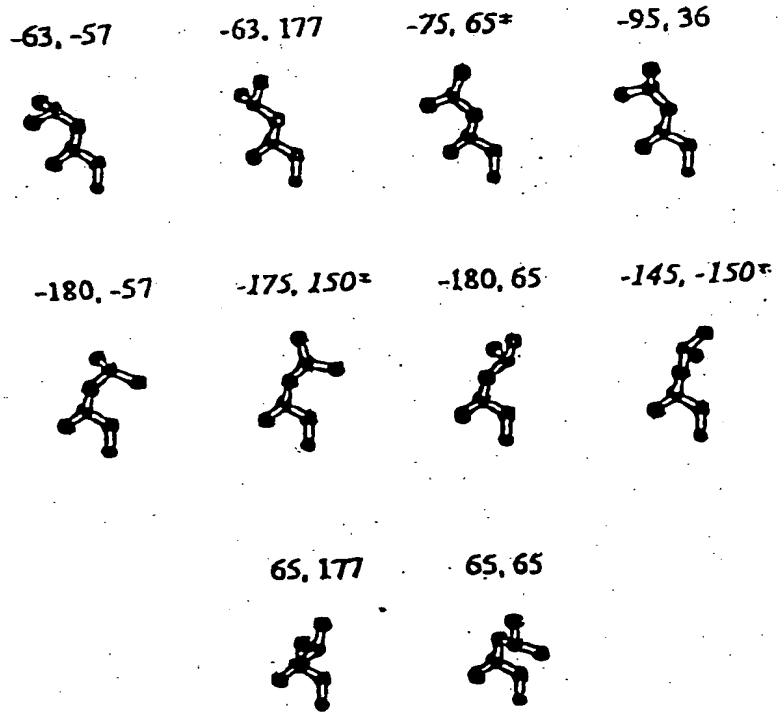


FIG. 12

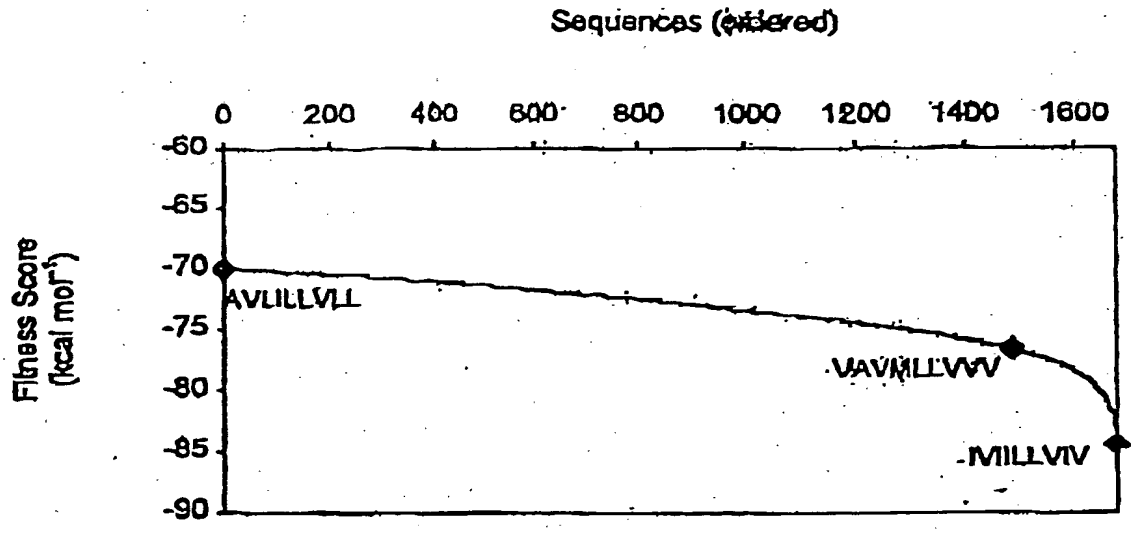

Sequences (ordered)
0
200
400
600
800
1000
1200
1400
1600
-60
-65
-70
-75
-80
-85
-90
Fitness Score (kcal mol-1)
AVLILLVLL
VAVMLLVVV
MILLVIV

FIG. 13

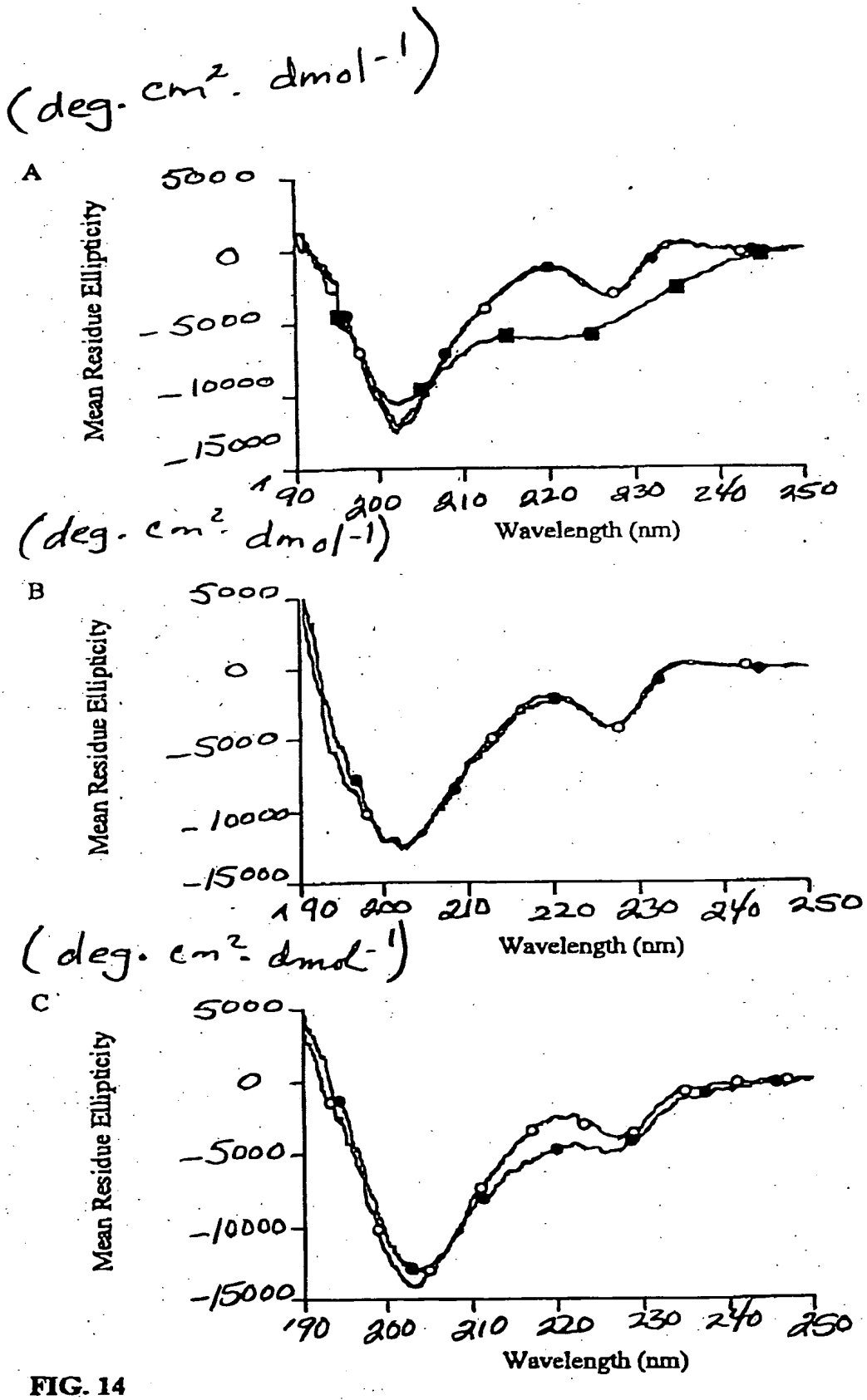

(deg. cm². dmol⁻¹)
A
Mean Residue Ellipticity
5000
0
-5000
-10000
-15000
190
200
210
220
230
240
250
Wavelength (nm)
(deg. cm². dmol⁻¹)
B
Mean Residue Ellipticity
5000
0
-5000
-10000
-15000
190
200
210
220
230
240
250
Wavelength (nm)
(deg. cm². dmol⁻¹)
C
Mean Residue Ellipticity
5000
0
-5000
-10000
-15000
190
200
210
220
230
240
250
Wavelength (nm)

FIG. 14

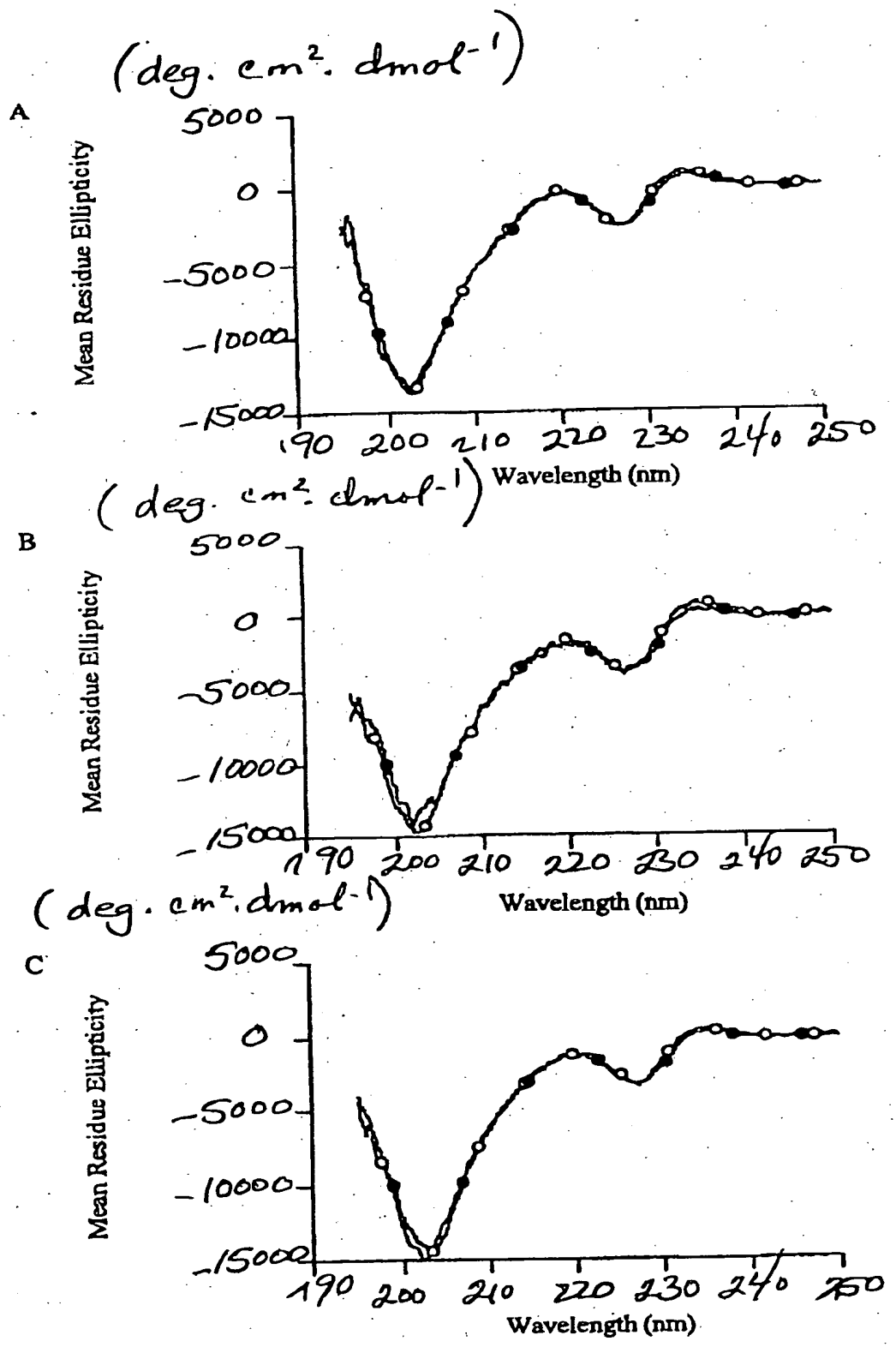

(deg. cm². dmol⁻¹)
A
Mean Residue Ellipticity
5000
0
-5000
-10000
-15000
190 200 210 220 230 240 250
Wavelength (nm)
(deg. cm². dmol⁻¹)
B
Mean Residue Ellipticity
5000
0
-5000
-10000
-15000
190 200 210 220 230 240 250
Wavelength (nm)
(deg. cm². dmol⁻¹)
C
Mean Residue Ellipticity
5000
0
-5000
-10000
-15000
190 200 210 220 230 240 250
Wavelength (nm)

FIG. 15

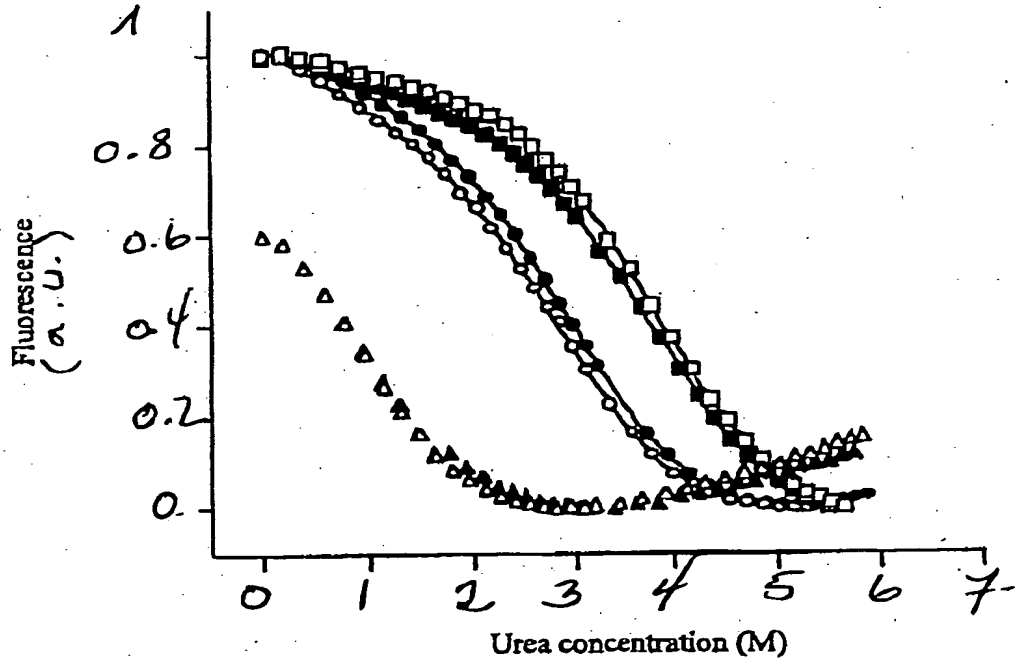


FIG. 16

(a.u. = arbitrary units)

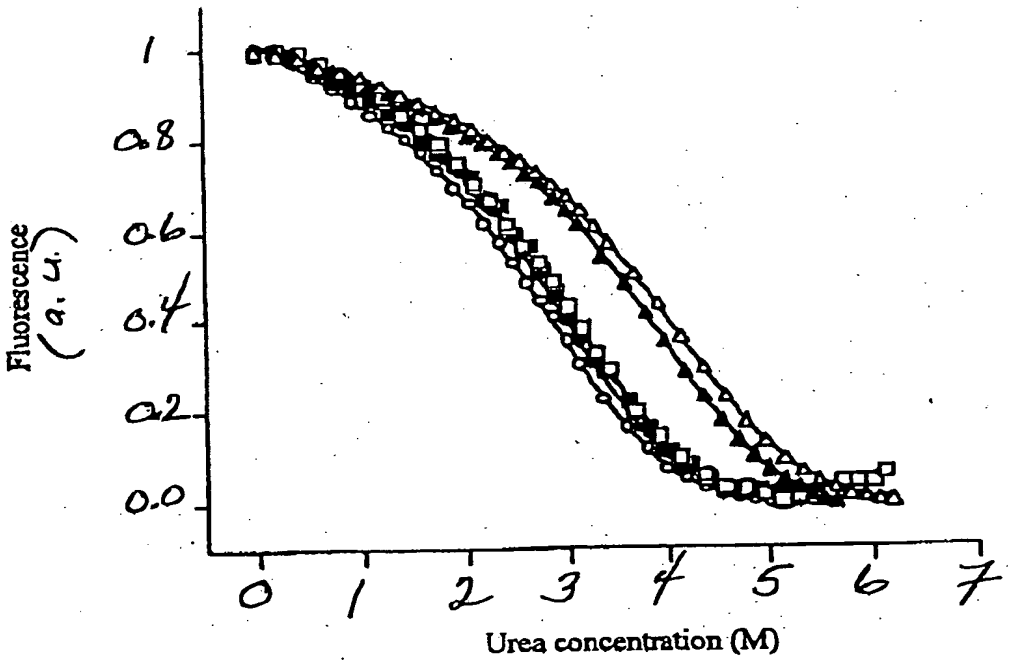


FIG. 17

(a.u. = arbitrary units)